Abstract

Emotion corpora are typically sampled based on keyword/hashtag search or by asking study participants to generate textual instances. In any case, these corpora are not uniform samples representing the entirety of a domain. We hypothesize that this practice of data acquisition leads to unrealistic correlations between overrepresented topics in these corpora that harm the generalizability of models. Such topic bias could lead to wrong predictions for instances like “I organized the service for my aunt’s funeral.” when funeral events are overrepresented for instances labeled with sadness, despite the emotion of pride being more appropriate here. In this paper, we study this topic bias both from the data and the modeling perspective. We first label a set of emotion corpora automatically via topic modeling and show that emotions in fact correlate with specific topics. Further, we see that emotion classifiers are confounded by such topics. Finally, we show that the established debiasing method of adversarial correction via gradient reversal mitigates the issue. Our work points out issues with existing emotion corpora and that more representative resources are required for fair evaluation of models predicting affective concepts from text.

1 Introduction

Emotion analysis is typically formulated as the task of emotion classification, i.e., assigning emotions to textual units such as news headlines, social media or blog posts. Emotion classification is applied across various domains, ranging from political debates (Mohammad et al., 2014) to dialogs (Li et al., 2017) and literary texts (Mohammad, 2011), and enable further use cases such as analyzing emotions of social media users (e.g., in response to the COVID-19 pandemic, Zhan et al., 2022), identifying abusive language using emotional cues (Safi Samghabadi et al., 2020) or developing empathetic dialog agents, e.g., for emotional support (Liu et al., 2021).

Emotions are thereby modeled as either discrete classes of basic emotions (Ekman, 1992; Plutchik, 2001), within the vector space of valence and arousal (Russell, 1980), or as the result of the emoter’s cognitive appraisal of the stimulus event (Scherer, 2005; Smith and Lazarus, 1990). Independent of which emotion theory is adopted, emotion data sets are commonly collected by searching for topics of interest, for instance with hashtags on social media (Schuff et al., 2017, i.a.) or by using specific subfora (Stranisci et al., 2022), in order to cover a variety of emotion labels instead of generally overrepresented ones. Another common approach is to ask study participants to report emotional episodes for a given emotion (Troiano et al., 2023, 2019; Scherer and Wallbott, 1994, i.a.). In that case, subjects are more likely to report important, long enduring, high-impact events than less relevant ones. Cases in which large corpora are uniformly sampled for annotation are comparably rare (Alm et al., 2005, i.a.).

We hypothesize that these established sampling procedures are harmful. They lead to topics overrepresented for specific emotions which allows the model to rely on spurious signals instead of actual emotion expressions. As an example, in “I enjoyed my birthday party;” a model might learn to associate the topic of “party” with joy, instead of inferring the emotion from the text (here, the verb). That might then lead to wrong predictions for texts such as “I did not like my party.”. We assume that this is also a reason for poor cross-corpus generalization of emotion classification (cf. Bostan and Klinger, 2018).

In this paper, we aim to understanding the prevalence and impact of this phenomenon in the context of emotion analysis. We answer the following research questions:

1. Are emotion datasets biased towards topics? We show that emotion datasets are biased towards topics, i.e., that there is a prototypical
association of topics with emotion labels specific for each corpus.

2. Is emotion classification influenced by topics?
Based on the observation of topic biases in datasets, we show that this bias also carries over to emotion prediction models.

3. Can the influence of topics on emotion classification be mitigated?
We show that the robustness of emotion classifiers can be improved by using established debiasing methods which reduce the impact of the topic bias on the classifiers.

We perform the experiments on emotion self-report corpora (Scherer and Wallbott, 1994; Troiano et al., 2023; Hofmann et al., 2020), social media data from Twitter (Schuff et al., 2017) and Reddit (Stranisci et al., 2022), as well as on fictional stories (Alm et al., 2005). With these annotated corpora, we cover (i) a variety of domains and (ii) multiple emotion models.

2 Related Work

2.1 Emotion Classification
Computational approaches to emotion analysis often adopt categories inspired by theories of basic emotions (Ekman, 1999; Plutchik, 1982), by modeling emotions as six (anger, fear, joy, sadness, disgust, surprise) or eight (adding anticipation, trust) discrete classes. Alternatives include the use of the valence–arousal vector space to position emotion categories (Russell, 1980) or focus on the aspect that emotions are caused by events that undergo a cognitive evaluation (Scherer, 2005; Smith and Lazarus, 1990). In the latter case, emotions are represented by appraisal variables, including, for instance, if the event requires attention, if the person involved is certain about what is happening, if the outcome requires further effort, is pleasant, or if the person has been responsible or can control the situation.

The emotion model is sometimes, but not always, chosen based on the domain a corpus stems from. For instance, Schuff et al. (2017) reannotate a stance detection corpus with Plutchik’s eight emotions due to their presumed universality. Alm et al. (2005) follow Ekman’s model for a similar reason. Scherer and Wallbott (1994); Hofmann et al. (2020) choose a set of self-directed emotions because their data consists of self-reports. Troiano et al. (2023) use a larger set of emotions, and also annotate appraisal dimensions because of the prevalence of event descriptions in the texts they collected, similarly to Stranisci et al. (2022).

To develop automatic emotion classification methods, as in many areas of NLP, transformer-based pre-trained language models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have been found to consistently outperform previous state-of-the-art approaches. These models are fine-tuned on domain-specific corpora. Bostan and Klinger (2018) show for 14 popular emotion datasets that a cross-corpus prediction performance is drastically lower than for in-corpus classification. We hypothesize that a major part of what makes a domain unique is the distribution of topics.

2.2 Bias
Bias has been found to affect various textual resources, including those to support hate-speech detection (Wich et al., 2020), sentiment analysis (Wang et al., 2021), machine translation (Stanovsky et al., 2019) or argument mining (Spliethöver and Wachsmuth, 2020). In general, the term bias refers to the phenomenon that machine learning models adopt latent, “non-generalizable features” (Shah et al., 2020) from the training data, such as domain-specific terms, contexts, or text styles. In consequence, the biased representation leads to erroneous results when applied to a domain where the alleged standard does not hold (cf. Hovy and Prabhumoye, 2021), which can lead to harmful impact on various groups in our society.

Topic bias originates in skewed topic representations. Wiegand et al. (2019), for instance, find the topic of soccer to be almost exclusively associated with abusive language, caused by the sampling procedure. In this paper, topic bias is understood to comprise two of these concepts: First, the association of certain emotion or appraisal labels with certain topics and second, the resulting bias in a classifier towards certain topics when predicting the emotion and appraisal labels.

Detection and Mitigation. For detecting bias contained within pre-trained models and word embeddings, Caliskan et al. (2017) introduce the Word Embedding Association Test (WEAT) and Kurita et al. (2019) investigate gender bias within BERT word embeddings. Wiegand et al. (2019) calculate the pointwise mutual information between words and abusive language annotations. Nejadgholi and Kiritchenko (2020) train a topic model on a dataset and perform a qualitative analysis of the result.
Bias mitigation is addressed at either the data or the modeling level. Wiegand et al. (2019) sample additional texts of the overrepresented class. Barikeri et al. (2021) augment training data by instance duplication, replacing the biased term with an inverse term. He et al. (2019) tackle the bias correction during training by developing an intentionally biased classifier in order to identify the features that exhibit bias. This information is then used to train a debiased classifier which compensates for the biased features. Qian et al. (2019) adapt the language model’s loss function in order to mitigate gender bias, introducing a new term to the loss function that aims at equalizing the probability of male and female words. In the context of mitigating the influence of domains on classification, gradient reversal has proven effective (Ganin et al., 2015).

3 Methods & Experimental Setting

We will now explain our method for topic-bias detection in emotion corpora and then the experimental setting to evaluate established mitigation methods in this domain.\footnote{The repository to replicate our experiments will be made available via \url{https://www.bamnlp.de/resources/}.

Definitions. We consider six different corpora, where each corpus $c \in C$ is modeled as a tuple consisting of a set of topic labels $T_c$, a set of instances $I_c$ and a set of annotation labels $L_c$, where $L_c$ is either from the set of overall appraisals ($L_c \subseteq A_C$) or emotion labels ($L_c \subseteq E_C$), where $A_C \cap E_C = \emptyset$.

Further, each instance $i_c \in I_c$ consists of a text $s_{i,c} = (s_1, s_2, \ldots, s_n)$, a topic label $t_{i,c} \in T_c$ and a set of emotion or appraisal labels $L_{i,c} = \{a_1, \ldots, a_k\} \subseteq L_c$. Some of the corpora we consider are labeled with multiple, i.e., one or more emotions. Appraisals are always annotated in a multi-label setting.

3.1 Topic-based Bias Detection

Inspired by Wiegand et al. (2019); Nejadgholi and Kiritchenko (2020), we train separate emotion classifiers tasked with predicting either the emotion or appraisal label $a \in L_{i,c}$, for each topic $t_{c}^{\text{out}} \in T_c$ in a given corpus. In the subset of the corpus used for training the classifier ($T_c^{\text{train}}$), instances with the topic label $t_{c}^{\text{out}}$ are excluded, i.e., $T_c^{\text{train}} = \{t_{i,c} | t_{i,c} \in T_c, t_{i,c} \neq t_{c}^{\text{out}}\}$. The number of classifiers trained for a given corpus $c$ is thus equal to $|T_c|$.

The classifiers are evaluated in two distinct settings: In the inTopic setting, multiple testsets are sampled from the corpus, one for each topic except $t_{c}^{\text{out}}$. Each testset is thus defined in relation to the respective held-out topic: $t_{c}^{\text{in}} = T_c \setminus \{t_{c}^{\text{out}}\}$. Thus, the union of all $t_{c}^{\text{in}}$ per corpus reflects $T_c^{\text{train}}$. Therefore, a classifier trained on $T_c^{\text{train}}$ is evaluated on all $t_{c}^{\text{in}}$ of corpus $c$. For the CrossTopic setting, the classifier is evaluated on the held-out topic $t_{c}^{\text{out}}$ which is not part of the training set $T_c^{\text{train}}$. In both settings, we calculate averages across folds which leads to a performance estimate whose comparisons are meaningful. Figure 1 visualizes this setup.

Topic Modeling. While emotion and appraisal annotations stem from the labels of the respective corpora, the topic labels need to be inferred from the data. We use BERTopic (Grootendorst, 2022), as it supports pre-trained transformer models to detect the semantic relations on sentence-level as well as HDBSCAN for clustering, averting the need of determining a fixed number of topics per dataset. This method has proven effective in previous research (Xu et al., 2022; Kellert and Mahmud Uz Zaman, 2022; Eklund and Forsman, 2022).

3.2 Bias Mitigation

We compare two established methods for debiasing the models with respect to topics.

Word Removal. As a straight-forward approach which still often shows a good performance (Dayanik and Padó, 2021, i.a.), the respective topic words are removed from the corpus. Specifically, we remove the most indicative words for each topic, according to the probabilities of the topic model.

Gradient Reversal. We compare this approach to the well-established method of adversarial learning through gradient reversal (Ganin et al., 2015). We extend the emotion/appraisal classifier by a topic predictor and gradient reversal layer, with the purpose of reversing the gradient (by multiplying it with $-\lambda$) of the following layer during back-propagation. Implementation details for all applied
### 3.3 Data

We consider six corpora, each annotated for emotions or appraisal dimensions. We use the ISEAR (Scherer and Wallbott, 1994), SSEC (Stance Sentiment Emotion Corpus; Schuff et al., 2017) and TALES (Alm et al., 2005) corpora for emotion analysis and the APPREDDIT corpus (Stranisci et al., 2022) for appraisal analysis. From the ENVENT (Troiano et al., 2023) and ENISEAR (Troiano et al., 2019) corpora we use both annotation layers.

The corpora differ in size, annotation setup and – most relevant for us – in the way the instances are sampled and which topics are covered: ISEAR and ENISEAR were created by asking study participants to report and describe events that caused a predefined emotion. ISEAR has been collected in an in-lab setup and ENISEAR via crowdsourcing. Since participants were free to report any event that elicited one of the given emotions, they were also free in their choice of topic. This procedure is in fact expected to create a topic bias, because more important topics cause more intense emotions and are therefore more likely to be recalled. Therefore, Troiano et al. (2019) add diversification method to the otherwise similar setup. They mention topics that the study participants shall not report on.

In the SSEC corpus, Schuff et al. (2017) re-annotate Twitter posts originally collected by Mohammad et al. (2016). The original purpose of the text collection was to study sentiment and stance. Therefore, they have been collected with specific hashtags corresponding to topics “Atheism”, “Climate Change is a Real Concern”, “Feminist Movement”, “Hillary Clinton”, and “Legalization of Abortion”. Arguably, we could have relied on these topics in the data, however for comparability in our experiments, we also use the topic modelling approach for this dataset.

The APPREDDIT corpus provides appraisal annotations of Reddit posts, sourced from subreddits mostly connotated with negative sentiment (Anger, offmychest, helpmecope anxiety, i.a.). The TALES corpus (Alm et al., 2005) features literary texts, specifically fairy tales by various authors. Here, sentences from uniformly sampled stories are the unit of annotation.

In order to enable inter-comparability, we map the varying annotation schemes onto a unified scheme. More information on the datasets is in Appendix B.

### 4 Results

We will now present the results to answer the research questions introduced in Section 1.

#### 4.1 Are emotions biased towards topics?

**Topic Modelling Results.** Table 1 reports the results of the topic modeling at the overall corpus level, including the number of topics, the average size (number of instances) and the list of topic labels \( L_s \) for each corpus. The topic labels are defined manually, based on the ten most representative words for each topic.

The size of topics, i.e., the number of instances associated with it, varies across corpora (see \( \varnothing \) and \( \text{STD} \)). The number of topics ranges from 8 (ENVENT) to 13 (ENISEAR), while ISEAR, TALES...
and APPREDDIT comprise 10, ISEAR 11 topics.

An important finding is that, despite not being informed in a supervised manner regarding the emotion labels, the topics reflect the individual corpus’ domain and sampling methods. ISEAR, enISEAR and enVENT, all of which are compiled by querying emotionally connotated event-descriptions, feature generic and everyday topics, e.g., love, dogs or driving. In SSEC the topic modeling corresponds to the keyword-based sampling based on the original intention to perform stance detection. In APPREDDIT, topics appear to be indicative of the subreddit they are sourced from. For instance, the topic of depression is related to the subreddit “mentalhealth”. The variety of relationship-related topics (romantic relationships, love, platonic relationships) reflects the various subreddits revolving around these topics, e.g., “relationship advice” or “Dear Ex” (cf. Stranisci et al., 2022 for the exhaustive list of sampled subreddits). The topics in TALES appear most varied. Some topics correspond to generic concepts within fairytales (birds, flowers, royalty), while others are representative of specific fairy tales\(^2\).

### Emotion–Topic Relation.

We will now look at the relation between emotions and topics from the dataset perspective. At first glance, such relations can already be observed in topics that revolve around specific emotions, such as shame, fear (both in ISEAR), anger (APPREDDIT) or, more general, feelings (enVENT). In order to assess whether these equivalences on the lexical level are also present in

\(^2\)The most representative terms for the topic labeled as Tabitha Twitchit comprise the names of fictional characters from the kids stories by Beatrix Potter. Further, the topic old english appears to be based on lexical features alone (e.g., “thou”, “thee”, “thy”).

the respective emotion annotations, we report the normalized pointwise mutual information between topics and their associated emotion annotations in Figures 2 and 3.\(^3\) For ISEAR (Figure 2), we observe that the topics of shame and fear are positively correlated with the emotion label of the same class. Further, emotionally correlated topics are death (with sadness), alcohol and animals (both disgust), accidents (fear) and exams with joy (all positive). Negative correlations can be observed for alcohol and joy, as well as for love and fear.

The observations for enVENT are similar (Fig. 3), with positive correlations between dogs and disgust as well as driving and fear. Although these are consistent with correlations of similar topics in

\(^3\)We focus our analysis on select datasets and report results for the remaining corpora in Appendix C.
**Table 2: Results for CROSSTOPIC and INTOPIC experiments and differences between them for all experimental series.**

For each experimental setup, we show results for the baseline without debiasing (BL) and for the two debiasing methods of word removal (WR) and gradient reversal (GR).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>CROSSTOPIC</th>
<th>INTOPIC</th>
<th>Δ INTOPIC-CROSSTOPIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BL</td>
<td>WR</td>
<td>GR</td>
</tr>
<tr>
<td>ISEAR</td>
<td>68</td>
<td>70</td>
<td>71</td>
</tr>
<tr>
<td>ENISEAR</td>
<td>69</td>
<td>54</td>
<td>68</td>
</tr>
<tr>
<td>SSEC</td>
<td>46</td>
<td>37</td>
<td>23</td>
</tr>
<tr>
<td>TALES</td>
<td>84</td>
<td>84</td>
<td>82</td>
</tr>
<tr>
<td>ENVENT</td>
<td>51</td>
<td>51</td>
<td>54</td>
</tr>
<tr>
<td>Average</td>
<td>62</td>
<td>57</td>
<td>58</td>
</tr>
</tbody>
</table>

ISEAR (animals and disgust, accidents and fear), the PMI in ENVENT are consistently lower.

The ENVENT offers itself to compare the emotion–topic and appraisal–topic correlations (Figure 4). The highest positive correlation is between covid and chance control, i.e., covid-related events are appraised as out of control by the emoter. The topic of covid is further (slightly) negatively correlated with self control (thus, the complement to chance control) and self responsibility. This direct comparison on ENVENT shows that the correlations between topics and appraisals are less distinct than for emotions.

### 4.2 Is emotion classification influenced by topics?

What arises from the observation that topics and emotions (and topics and appraisals) are indeed correlated is the question whether this relation is reflected in classifiers. To this end, Table 2 shows results for CROSSTOPIC and INTOPIC experiments.

Following the assumption that emotion and appraisal classifiers are biased towards topics, the INTOPIC setting is hypothesized to score higher than the CROSSTOPIC setting. The difference between these two settings is shown in the Δ INTOPIC-CROSSTOPIC—BL column. Across all corpora, we see that all INTOPIC scores are higher than the CROSSTOPIC scores—the Δ is positive but varies: The highest discrepancy is observed for ISEAR (+9), while it is negligible for SSEC, TALES and ENVENT (in the appraisal classification setting) and APPREDIT (+2). In comparison, ENVENT (for emotion classification) as well as emotion and appraisal classification on ENISEAR show moderate improvement when evaluated IN-TOPI

### 4.3 Can the influence of topics on emotion classification be mitigated?

To understand if the discrepancy between the CROSSTOPIC and INTOPIC results can be mitigated with debiasing methods, we show the results also in Table 2 (columns WR for word removal and GR for gradient reversal). The large Δ value reported for ISEAR in Table 2 leads to the diagonal values (CROSSTOPIC) in Figure 5 to be lower than the average of all other results of the same held-out topic (INTOPIC). However, the CROSSTOPIC scores are still comparably high. Particularly interesting is the topic of death. When this is absent from the training data, the classifier performs much worse on all testsets, both INTOPIC and CROSSTOPIC. Analogously, the topic fear appears to contain instances easier to classify, no matter which held-out topic is absent from the training data. The only exception is the mentioned topic death, and, although to a lesser extent, the CROSSTOPIC setting of the topic fear.
The direct comparison shows that the substantial difference in performance between the mitigation methods. In the aforementioned corpora, an improvement can only be observed in the GR-setting. When WR is applied, ISEAR and ENISEAR even show a decrease in performance. While the SSEC corpus would also have the potential to be improved with the method, the classifier relied too substantially on the topic information and cannot find enough signal for emotion classification such that the method may work.

For the appraisal prediction, we also observe an improvement for event-centered corpora ENISEAR and APPREDIT, but not for ENVENT. Throughout all experiments, we observe that topic information removal is disadvantageous for appraisal prediction. We take this as an indicator that the classifiers indeed find information on the emotion expression outside of topic information. However, the appraisal information needs to be inferred from the topic of the text and cannot be found elsewhere.

5 Analysis

To provide an intuition how the predictions of the model changes with the topic mitigation, we show examples in Table 3. For each example sentence we see the corresponding topic label (according to the topic model), the gold emotion annotation and the CrossTopic-predictions with (WR, GR) or
Table 3: Example predictions for instances from the ISEAR corpus, including assigned topic and gold emotion label. Predictions are reported for the CrossTopic-setting (trained on all instances except those labeled with respective topic in column Topic) when applying no mitigation method (BL), word removal (WR) and gradient reversal (GR). Predictions in bold represent correspondence with gold label.

<table>
<thead>
<tr>
<th>ID</th>
<th>Text</th>
<th>Topic</th>
<th>Gold</th>
<th>BL</th>
<th>Wr</th>
<th>GR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>When one of my closest friends died unexpectedly</td>
<td>death</td>
<td>sadness</td>
<td>joy</td>
<td>disgust</td>
<td>sadness</td>
</tr>
<tr>
<td>2</td>
<td>When my uncle comes (3 times a year) for the traditional Christmas dinner with my grandparents and other relatives and is very drunk.</td>
<td>alcohol</td>
<td>disgust</td>
<td>anger</td>
<td>shame</td>
<td>disgust</td>
</tr>
<tr>
<td>3</td>
<td>When my fiancee travelled 2000 Km to visit me, and I hadn’t seen her for 4 months.</td>
<td>love</td>
<td>joy</td>
<td>sadness</td>
<td>sadness</td>
<td>joy</td>
</tr>
<tr>
<td>4</td>
<td>Passing an exam I did not expect to pass.</td>
<td>exam</td>
<td>joy</td>
<td>fear</td>
<td>fear</td>
<td>fear</td>
</tr>
<tr>
<td>5</td>
<td>When I was admitted to a certain school as a student.</td>
<td>exam</td>
<td>love</td>
<td>sadness</td>
<td>sadness</td>
<td>fear</td>
</tr>
</tbody>
</table>

Without (BL) applying de-biasing methods.

Example 1 is assigned the topic death and is annotated with sadness. With no mitigation method applied, a CrossTopic-classifier (i.e., which has not seen any sentences belonging to the topic death during training) falsely predicts joy (BL). We hypothesize that the erroneous classification is due to a bias towards the topic of love (which is correlated with joy), represented by the term “friends”. If word removal is applied, a different but equally incorrect label is predicted (disgust). Apparently, removing any words associated to topics from the input does mitigate the bias observed in the BL prediction, but removes too much information. However, when using gradient reversal, the bias is mitigated and the correct label sadness is predicted. Similar cases can be observed in Examples 2, 3, 5 and 6.

Example 4 shows a different pattern. Despite achieving de-biasing in the above cases, there are also examples where gradient reversal fails to mitigate the bias and predict the correct emotion label. None of the two mitigation methods leads to a correct prediction. Instead, all CrossTopic-classifiers assign fear. Presumably, this is because of the phrase “did not expect” which expresses a future-directed, misalignment with the predictability of events. This aspect might in itself be another possible form of appraisal bias.

6 Conclusion

We based our study on the observation that emotion analysis corpora are commonly sampled based on keywords or following other methods that are risky to lead to distributions that are not representative for the entirety of a domain. We contributed a better understanding how far this issue can be found in emotion corpora and if models fine-tuned on them rely on such spurious signals.

The analysis of topic distributions in emotion corpora yields that they are, indeed, biased towards topics. The degree of bias varies: Some corpora exhibit prototypical topics for certain emotions, while in others, only weak correlations between topic and emotion distribution can be observed. We hypothesize this is because of the respective sampling strategies: If the sampling method is biased, i.e., if certain topics are over-represented for a given emotion, topic bias emerges.

In the cases in which topic and emotion distributions are highly correlated, this topic bias is also found to be reflected in the resulting classifier. For mitigating this bias in emotion classifiers, gradient reversal proved to be useful. It allows the classifier to make use of available topic information without relying solely on it for making the classification decision.

Our results suggest that classifiers in which the topic bias is mitigated may have a higher performance across corpora, yet, this needs to be evaluated in future work. Further, we assume that prompt-learning or other few-shot modeling methods might suffer less from topic biases in corpora. If this is true, this opens a new research direction of selecting non-bias-inducing instances for emotion and appraisal classification.

Finally, the difference between topic–emotion and topic–appraisal correlations requires further analysis. We hypothesize that this is because appraisals are more closely related to events than general emotion labels.
Acknowledgements
This research has been conducted in the context of the CEAT project, KL 2869/1-2, funded by the German Research Foundation (DFG).

Limitations
We presented the first study on topics as unwanted confounders for emotion analysis. We focused on a set of popular corpora, but cannot make any judgments regarding corpora that we did not study. We are confident that similar effects can be found in other resources, but this still needs to be analyzed.

Another limitation is the pragmatic decision that the contextualized embeddings used by our emotion/appraisal predictors and the topic modeler are not the same. The representations used for topic clustering are provided by sentence-transformer models, while we leverage RoBERTa embeddings for emotion and appraisal classification. This potentially introduces an uncontrolled variance in our experiments. Using identical embedding models for both steps – or, alternatively, a joint embedding space – might reduce that variance and thus improve interpretability of the results.

Ethical Considerations
In our work, we do not develop or annotate corpora. We further do not collect data or propose new NLP tasks. Therefore, our work does not contribute potential biases originating from annotator or data selection. Instead, our goal is to understand biases better and contribute to a more fair emotion classification. We do not investigate how topic bias might cause harm in downstream applications.

Still, our topic analysis might be limited, for instance by the topic modeler chosen for the analysis and by the datasets that we studied. In real-world data applications, another topic modeling approach might be required. It is important to note that we do not make any statements which topics might have a negative impact on members of a society.

In general, emotion classifiers have a high potential to cause harm by making wrong predictions. Until the performance is on a higher, more reliable level and the effects of biases and other confounding variables are better understood, they should always be applied with caution. We propose that the analyses acquired with automatic emotion analysis methods should never be related to individuals. Instead, analysis should only be performed on an aggregated level.

References


A Implementation Details

Emotion/Appraisal Classifier. Following state-of-the-art approaches to emotion and appraisal classification (Demszky et al., 2020 Troiano et al., 2019), we fine-tune RoBERTa (Liu et al., 2019) as implemented in the Huggingface library (Wolf et al., 2020) on each corpus. For the classification, the output from the transformer layers is pooled and passed through a fully-connected dense layer (768 units). We apply ReLU activation (Agarap, 2019) and a dropout of 0.5 and a consecutive classification layer using softmax activation and binary cross-entropy loss for single-class classification (for ISEAR, TALES, and emotions in ENVENT). For the multi-class classification task (SSEC, APPREDDIT, ENISEAR and appraisals in ENVENT), we apply a sigmoid activation and categorical cross-entropy loss instead. The learning rate is set to $5 \times 10^{-5}$ across all experiments; the batch size is 16. We train each classifier for a maximum of 5 epochs and apply early stopping based on the validation accuracy (stops after two consecutive epochs without improvement). As optimizer, AdamW (Loshchilov and Hutter, 2019) is applied, weight decay is set to $10^{-5}$. Results are averaged over three different runs for each classification task.

Topic Modeling. BERTOPIC consists of a pipeline of components for features representation, dimensionality reduction, clustering and topic. We use a pre-trained sentence embedding (all-MiniLM-L6-v2, as implemented in Huggingface) for feature extraction, Accelerated Hierarchical Density Clustering (HDBSCAN; McInnes and Healy, 2017) as a clustering method, Uniform Manifold Approximation (UMAP; McInnes et al., 2020) for dimensionality reduction and tf-idf for retrieving the topics within the clusters. Although HDBSCAN does not require a pre-determined number of topics, it can be tuned by setting hyperparameters for the minimum cluster size and controlling the amount of outliers allowed within a cluster. We adapt these hyperparameters to each corpus individually, depending on its size.

Word Removal. The list of topic words to be removed in each corpus consists of the ten most representative words of each topic within the dataset. The most representative words, i.e., the top $k$ words per topic are determined by the probability that BERTOPIC assigns to each word, i.e., the word’s probability to be assigned a certain topic label. Therefore, $k$ is a hyperparameter determining the trade-off between general classification performance and topic-influence: Increasing $k$ increases the potential impact of the de-biasing method (as less topic-specific features are available to the classifier), but, at the same time, decreases the general classification as less and less features are available overall. Further, by choosing a higher $k$, more words which are less representative for a given topic are removed as well, thus introducing noise to the experiment. Here, $k$ is set to 10. Setting $k = 3$ or $k = 5$ were considered as well, but did not show a considerable change in performance compared to the non-mitigated baseline classifier (BL). This hyperparameter choice is further supported by the observation that the top $k$ representative words often comprise variations of the same word or concept. For example, in ISEAR, the ten most representative words for the topic theft consist of “theft”, “stealing”, “stole”, “thief”, “robbery”, “thieves”, “stolen”, “borrowed”, “robbers” and “cash”. A higher $k$ thus covers a broader range of morphological (“stealing”, “stole”, “stolen” and “thief”), as well as semantic (”theft”, ”robbery”) variation. The chosen topic words are not removed from the input, but substituted with “...”. The number of masked topic words per corpus is summarized in Table 4.

Gradient Reversal. The gradient reversal layer (GRL) is implemented as described by Ganin et al. (2015), with the purpose of reversing the gradient (by multiplying it with $-\lambda$) of the following layer during backpropagation. Since the layer has no trainable (nor non-trainable) weights associated with it, the GRL has no effect during a forward pass and acts as an identity transformation. For the INOPIC-GR and CROSSOPIC-GR experiments conducted here, the GRL is added into the standard classifier architecture described above. The emotion classifier is coupled with an additional topic classification layer, equivalent to the single-class emotion classification layer, with the task of pre-

### Table 4: Number (#) of topics and the resulting number of removed (i.e., masked) topic words.

<table>
<thead>
<tr>
<th>Corpus</th>
<th># topics</th>
<th># masked topic words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISEAR</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>SSEC</td>
<td>11</td>
<td>110</td>
</tr>
<tr>
<td>TALES</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>ENVENT</td>
<td>8</td>
<td>80</td>
</tr>
<tr>
<td>APPREDDIT</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>ENISEAR</td>
<td>13</td>
<td>130</td>
</tr>
</tbody>
</table>
dicting the correct topic label $t_{i,c}$ for each instance. The topic classifier is connected via the GRL to the remaining layers of the network, i.e., the pre-trained RoBERTa model as well as the single dense layer. Since the gradient is reversed, all weights in the shared layer associated with the topic prediction task are decreased. A key factor in the implementation is the choice of $\lambda$ as it regulates the impact of the GRL. Again, choosing $\lambda$ is a trade-off between overall classification performance and de-biasing potency. To determine an optimal value for $\lambda$, standard emotion (or appraisal classifiers) are trained on each individual corpus for $\lambda$ values of 0.1, 0.3, 0.5, 1 and 3. Across corpora, a significant decrease in performance can be observed for any $\lambda > 0.1$. Therefore, $\lambda$ is set to 0.1 for all gradient reversal experiments.

B Data

Besides for their widespread use, the corpora are specifically selected for their variety in domain and text style. As bias in general and topic bias in particular is closely related to the respective dataset’s domain, annotation and sampling methods of a dataset, the following overview puts emphasis on these aspects. We provide a detailed description of the datasets used in this investigation, emphasizing on each dataset’s domain, annotation and sampling method. General corpus statistics are further provided in Table 6.

B.1 Corpora

ISEAR. The ISEAR corpus (Scherer and Wallbott, 1994) consists of 7,665 sentences which were sampled in an in-lab setting: Participants were presented with an emotion label and asked to report an event that elicited that particular emotion in them. Each event description is labeled with a single emotion from a set of eight (Ekman’s basic emotions plus shame and guilt). Since participants were free to report any event that elicited one of the given emotions, they were also free in their choice of topic. However, since participants were asked to report events specific to certain emotions, sample bias could have been introduced to the corpus (under the assumption that there are prototypical events for certain emotions).

ENISEAR. The corpus consist of 1001 event descriptions that were originally compiled by (Troiano et al., 2019) as a complement to ISEAR. The event descriptions were sampled analogous to ISEAR, but in a crowd-sourcing setup ( annotated for joy, sadness, anger, fear, disgust, shame and guilt). Here, ENISEAR also refers to the appraisal annotations which were added to the corpus by Hofmann et al. (2020): Attention, certainty, effort, pleasantness, responsibility and control. These additional annotations were provided by expert annotators.

SSEC. The Stance Sentiment Emotion Corpus (Schuff et al., 2017) consists of 4,868 Twitter posts. The original data stems from Mohammad et al. (2016) which Schuff et al. (2017) re-annotate for Plutchik’s eight basic emotions. The annotations are conducted by trained expert annotators. Since the original dataset by Mohammad et al. (2016) was developed for stance detection, the instances were sampled using keywords (i.e., hashtags) that contain a particular stance in favor (e.g., “#Hillary4President”) or against an entity (“#HillNo”). This type of keyword-based data sampling has been found to exhibit topic bias in related studies, e.g., on datasets of abusive language (Wiegand et al., 2019).

TALES. The TALES corpus (Alm et al., 2005) features 15,302 sentences from different fairytales. Sentences are labeled by experts with one of Ekman’s basic emotions (surprise is split into negative and positive surprise). Emotions are annotated from the perspective of the respective character.

crowd-ENVENT. Analogous to ENISEAR, the crowd-ENVENT corpus (Troiano et al., 2023) consists of 6600 crowd-sourced, self-reported event descriptions. Each description is annotated for 21 appraisal dimensions\(^4\), each rated on a scale between 1 and 5, as well as for emotions (Ekman’s 6 basic emotions, plus shame, pride, boredom, relief, trust, shame, guilt and no emotion). Participants were free in their choice of topic, but the priming with an emotion label might influence the topic distribution (see ISEAR). In order to avoid oversampling descriptions of prototypical events, Troiano et al. apply a diversification method to foster more diverse event descriptions. The corpus additionally features crowd-sourced re-annotations of the event descriptions to investigate differences between the

\(^4\)Suddeness, familiarity, event predictability, pleasantness, unpleasantness, goal relevance, own responsibility, others’ responsibility, situational responsibility, expectation of consequences, goal support, urgency, own control, others’ control, situational control, acceptance of consequences, clash with internal standards and ideals, violation of (external) norms and laws, not consider, attention, effort.
reader’s and writer’s assessment of emotions and appraisals. However, these are not used here.

**APPREDIT.** The APPREDIT corpus (Stranisci et al., 2022) is annotated with appraisal dimensions. It comprises 780 Reddit posts, where each post contains at least one event description (1,091 events overall). The five appraisal labels (certainty, consistency, control, unexpectedness, responsibility) are based on (Roseman, 1991) and annotated by experts. The posts are sampled exclusively from a limited set of subreddits, mostly connotated with negative sentiment (Anger, offmychest, helpmecope anxiety, i.a.). This sampling procedure might introduce bias to the dataset.

### B.2 Aggregated Annotation Scheme

As depicted above, the corpora differ in their annotation schemes. In order to provide a more comparable analysis, the individual annotations are mapped onto an inter-corpora annotation scheme. For emotions, anger, disgust, fear, joy, sadness, surprise, no emotion and other are considered. This subset of emotion labels is based on basic emotions (Ekman, 1999). Beyond Ekman’s six emotions, the list accounts for other labels that frequently occur (see Table 6 for an overview). The same procedure is applied to appraisal labels. However, approaches to appraisal classification are even more diverse in annotation than emotion datasets. To account for this variation, the inter-corpora label set consists of 11 appraisal dimensions (suddenness, pleasantness, self control, chance control, self responsibility, other responsibility, goal support, predict consequences, attention, effort), however, only a subset of six labels is shared across two of the three corpora annotated with appraisals, while only two labels can be mapped to all three corpora (summarized in Table 7).

### C Other Emotion–Topic Relations

Figures 7 and 8 show the results for topic–emotion associations for the TALES and the SSEC corpora, analogously to the other resources in Section 4.
<table>
<thead>
<tr>
<th>Corpus</th>
<th>A</th>
<th>D</th>
<th>F</th>
<th>J</th>
<th>Sa</th>
<th>Sh</th>
<th>Su</th>
<th>No</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENVENT</td>
<td>550</td>
<td>550</td>
<td>550</td>
<td>550</td>
<td>550*</td>
<td>550*</td>
<td>2,200*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISEAR</td>
<td>1,096</td>
<td>1,096</td>
<td>1,095</td>
<td>1,094</td>
<td>1,096</td>
<td>2,189*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ENISEAR</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>143</td>
<td>286*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSEC</td>
<td>1388</td>
<td>440</td>
<td>274</td>
<td>815</td>
<td>414</td>
<td>177</td>
<td>1552</td>
<td>1077*</td>
<td></td>
</tr>
<tr>
<td>TALESEX</td>
<td>302</td>
<td>40</td>
<td>251</td>
<td>579</td>
<td>340</td>
<td>144</td>
<td>8,683</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Number of instances of each emotion class (after mapping; the asterisk (*) indicates that this class includes mapped labels, i.e., combining multiple classes into one aggregated, but not simple one-to-one mapping of equivalent labels (happiness $\rightarrow$ joy)).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Attention</th>
<th>Pleasantness</th>
<th>Suddenness</th>
<th>Self Control</th>
<th>Chance Control</th>
<th>Self Responsibility</th>
<th>Other Responsibility</th>
<th>Predict Consequences</th>
<th>Goal Support</th>
<th>Effort</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPREDICT</td>
<td>–</td>
<td>–</td>
<td>307</td>
<td>307</td>
<td>–</td>
<td>400</td>
<td>457</td>
<td>748</td>
<td>312</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ENVENT</td>
<td>4125</td>
<td>2261</td>
<td>3128</td>
<td>2142</td>
<td>1514</td>
<td>2597</td>
<td>3396</td>
<td>2841</td>
<td>2281</td>
<td>3210</td>
<td>6527*</td>
</tr>
<tr>
<td>ENISEAR</td>
<td>673</td>
<td>149</td>
<td>–</td>
<td>228</td>
<td>240</td>
<td>377</td>
<td>–</td>
<td>761</td>
<td>400</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 7: Number of instances of each appraisal class (after mapping; the asterisk (*) indicates that this class includes mapped labels, either by simple one-to-one mapping (happiness $\rightarrow$ joy), or by combining multiple classes into one aggregated).