Topic Bias in Emotion Classification

Maximilian Wegge[1] and **Roman Klinger**[1,2]
[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[2]Fundamentals of Natural Language Processing, University of Bamberg, Germany
maximilian.wegge@ims.uni-stuttgart.de
roman.klinger@uni-bamberg.de

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 at 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.[1]

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 $A_C \cap E_C = \varnothing$.

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 $L_{i,c} = \{a_j, \ldots, a_k\} \subseteq L_c$. 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^{train}), instances with the topic label t_c^{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 $|T_c|$.

[1]The repository to replicate our experiments will be made available via `https://www.bamnlp.de/resources/`.

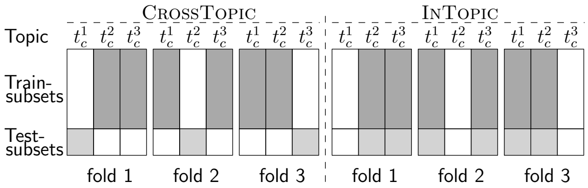

Figure 1: Visualization of the experimental setting for InTopic and CrossTopic predictions.

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^{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^{in} per corpus reflects T_c^{train}. Therefore, a classifier trained on T_c^{train} is evaluated on all t_c^{in} of corpus c. For the CrossTopic setting, the classifier is evaluated on the held-out topic t_c^{out} which is not part of the training set T_c^{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 backpropagation. Implementation details for all applied

	# Topics	∅ Topic	StD	Topic labels	# Instances	*Outlier*
ISEAR	10	525	290	love, exams, death, shame, school, animals, alcohol, accidents, fear, theft	7666	2412
SSEC	11	305	219	feminism, prayer, abortion, climate, clinton, twitter, trump, gay marriage, latino, swearing, patriotism	4870	1513
Tales	10	388	183	birds, flowers, tabitha twitchit, old english, piggies, royalty, dressmaking, hansel & gretel, boats, predators	10339	6457
enVENT	8	584	298	feelings, promotion, relationships, covid, dogs, graduation, pregnancy, driving	6600	1925
APPReddit	10	43	12	depression, everyday life, driving, love, romantic relationships, reddit, anger, death, platonic relationships, vaccination	780	352
enISEAR	13	58	25	death, dogs, accidents, theft, birth, food, affairs, UK politics, christmas, bullying, work, relationships, spooky	1001	245

Table 1: Number (#), average size (∅), standard deviation (StD), and manually assigned labels of the topics found by BERTopic for all corpora. All numbers exclude the outlier topic, whose number of instances is provided in the last column (*Outlier*). The topic labels are sorted by size, in decreasing order. The second to last column reports the number of all instances per corpus for reference. (We abbreviate the crowd-enVENT corpus in this paper as enVENT.)

methods are provided in Appendix A.

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_c) 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 ∅ and StD). The number of topics ranges from 8 (enVENT) to 13 (enISEAR), while ISEAR, Tales

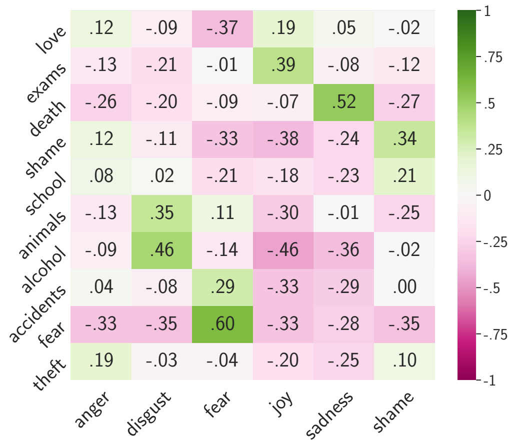

Figure 2: Normalized pointwise mutual information between topics and emotion annotations in ISEAR.

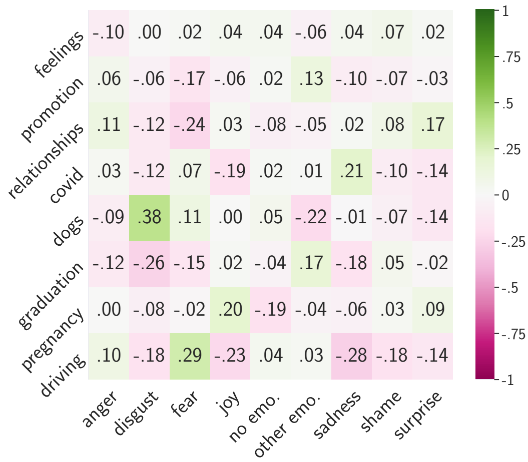

Figure 3: Normalized pointwise mutual information between topics and emotion annotations in ENVENT.

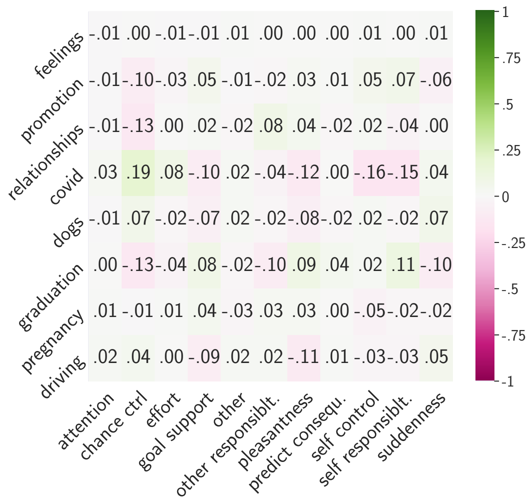

Figure 4: Normalized pointwise mutual information between topics and appraisal annotations in ENVENT.

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 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 (Fig. 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

[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").

[3]We focus our analysis on select datasets and report results for the remaining corpora in Appendix C.

		CrossTopic					InTopic					$\Delta^{\text{InTopic}}_{\text{CrossTopic}}$		
	Corpus	Bl	Wr	Gr	$\Delta^{\text{Bl}}_{\text{Wr}}$	$\Delta^{\text{Bl}}_{\text{Gr}}$	Bl	Wr	Gr	$\Delta^{\text{Bl}}_{\text{Wr}}$	$\Delta^{\text{Bl}}_{\text{Gr}}$	Bl	Wr	Gr
Emotion	ISEAR	59	59	65	0	6	68	70	71	2	3	9	11	6
	enISEAR	69	54	68	−15	−1	74	69	72	−5	−2	5	15	4
	SSEC	46	37	23	−12	−23	47	39	25	−8	−22	1	2	2
	Tales	84	84	82	0	−2	85	85	83	0	−2	1	1	1
	enVENT	51	51	54	0	3	55	55	57	0	2	4	4	3
	Average	62	57	58	−5	−4	66	64	62	−2	−4	4	7	4
Appraisal	enISEAR	70	56	54	−14	−16	75	57	56	−18	−19	5	1	2
	enVENT	63	61	44	−2	−19	64	61	45	−3	−19	1	0	1
	APPReddit	66	56	56	−10	−10	68	55	56	−13	−12	2	−1	0
	Average	66	57	51	−9	−15	69	57	52	−12	−17	3	0	1

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).

ISEAR (*animals* and *disgust*, *accidents* and *fear*), the PMI values 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 $\Delta^{\text{InTopic}}_{\text{CrossTopic}}$–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 neglectable for SSEC, Tales and enVENT (in the appraisal classification setting) and APPReddit (+2). In comparison, enVENT (for emotion classification) as well as emotion and appraisal classification on enISEAR show moderate improvement when evaluated InTopic (+4, +5, +5, respectively). Overall, the Δ values are similar (on average) between emotion and appraisal classification.

These results show that the topic influences the predictions negatively, but does not allow any insight if these results mostly stem from one emotion label or are the same across labels. To analyze this aspect, Figure 5 reports the F_1-scores obtained on each topic-specific subset for each held-out topic. The diagonal thus depicts the CrossTopic setting. All other cells correspond to the InTopic setting.

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*.

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).

Do the mitigation methods lower the performance for each setting separately or do they im-

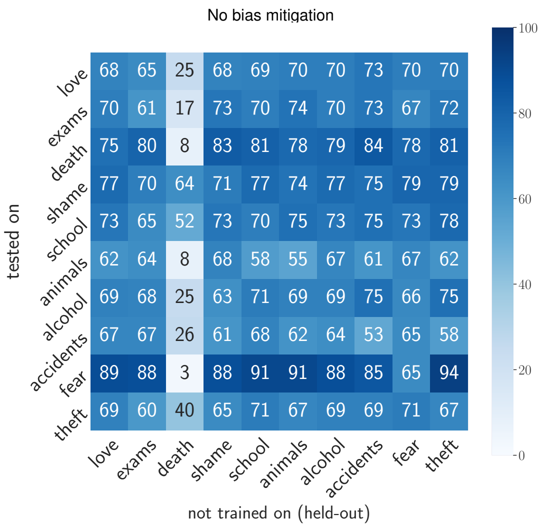

Figure 5: Micro-average F_1 for each topic-specific test set in ISEAR, for each held-out topic (CROSSTOPIC/INTOPIC). No mitigation method is used (BL setting).

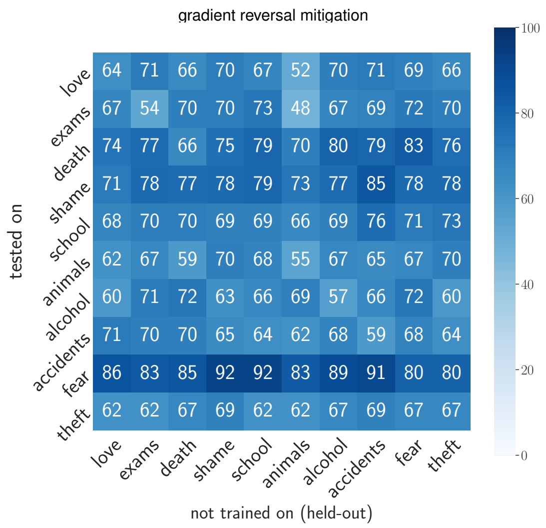

Figure 6: Micro-average F_1 for each topic-specific test set in ISEAR, for each held-out topic (CROSSTOPIC/INTOPIC). Gradient reversal is used as a mitigation method (GR setting).

prove it? The answer can be found in the $\Delta^{\text{BL}}_{\text{WR}}$ $\Delta^{\text{BL}}_{\text{GR}}$ columns. In the INTOPIC setting, most of these values are negative – the mitigation method removes information helpful for emotion classification. The only exception is the ISEAR corpus for emotion classification, where the method in fact improves the result. The negative difference is most pronounced for SSEC and nearly negligible for the other corpora for emotion classification. The results carry over to the CROSSTOPIC setting: For SSEC, emotion classification performance is substantially lower, while the difference is neglectable for most other corpora. Only ENISEAR (for emotion classification) shows a similarly significant drop in performance when WR is applied. For ISEAR, however, the emotion classification is improved. To provide more detail on where this CROSSTOPIC improvement takes place, we compare the detailed INTOPIC/CROSSTOPIC results for the BL- and GR-settings in Figures 5 and 6, respectively. The direct comparison shows that the substantial impact of the topic *death* on CROSSTOPIC emotion classification (Figure 5) is mitigated when applying the GR-mitigation method (6).

Do the mitigation methods lower the performance discrepancy between the INTOPIC *and* CROSSTOPIC *predictions?* To find the answer to this question, we compare the delta values BL–WR and BL–GR at the right of Table 2 ($\Delta^{\text{INTOPIC}}_{\text{CROSSTOPIC}}$). A lower delta value for the mitigation method than for the BL is an indicator that the method improves the classifier. In the emotion classification setup, this is the case for ISEAR and, to a lower extend, for ENISEAR and ENVENT. These are the corpora that are particularly designed to include event descriptions. However, there is a 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 APPREDDIT, 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

ID	Text	Topic	Gold	CROSSTOPIC BL	CROSSTOPIC WR	CROSSTOPIC GR
1	When one of my closest friends died unexpectantly	death	sadness	joy	disgust	**sadness**
2	When my uncle comes (3 times a year) for the traditional Christmas dinner with my grandparents and other relatives and is very drunk.	alcohol	disgust	anger	shame	**disgust**
3	When my fiancee travelled 2000 Km to visit me, and I hadn't seen her for 4 months.	love	joy	sadness	sadness	**joy**
4	Passing an exam I did not expect to pass.	exam	joy	fear	fear	fear
5	When I was admitted to a certain school as a student.	exam	joy	shame	shame	**joy**
6	Unexpected visit by a close friend, whom I hadn't seen for half a year.	love	sadness	sadness	fear	**sadness**

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.

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 judgements 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

Abien Fred Agarap. 2019. Deep learning using rectified linear units (relu).

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Erenay Dayanik and Sebastian Padó. 2021. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, Online. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anton Eklund and Mona Forsman. 2022. Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages

635–643, Abu Dhabi, UAE. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Paul Ekman. 1999. *Basic Emotions*, chapter 3. John Wiley & Sons, Ltd.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2015. Domain-adversarial training of neural networks.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Olga Kellert and Md Mahmud Uz Zaman. 2022. Using neural topic models to track context shifts of words: a case study of COVID-related terms before and after the lockdown in April 2020. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 131–139, Dublin, Ireland. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif Mohammad, Xiaodan Zhu, and Joel Martin. 2014. Semantic role labeling of emotions in tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, Baltimore, Maryland. Association for Computational Linguistics.

Isar Nejadgholi and Svetlana Kiritchenko. 2020. On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online. Association for Computational Linguistics.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Robert Plutchik. 2001. The Nature of Emotions. *American Scientist*, 89(4):344.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Ira J. Roseman. 1991. Appraisal determinants of discrete emotions. *Cognition & Emotion*, 5:161–200.

James Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. 2020. Attending the emotions to detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88, Online. Association for Computational Linguistics.

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

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Craig A. Smith and Richard S. Lazarus. 1990. Emotion and adaptation. *Handbook of Personality: Theory and Research*, pages 609–637.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. 2022. APPReddit: a corpus of Reddit posts annotated for appraisal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France. European Language Resources Association.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, 49(1):1–72.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, and Yi Chang. 2021. Eliminating sentiment bias for aspect-level sentiment classification with unsupervised opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3002–3012, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. 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.

Xiao Xu, Gert Stulp, Antal Van Den Bosch, and Anne Gauthier. 2022. Understanding narratives from demographic survey data: a comparative study with multiple neural topic models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 33–38, Abu Dhabi, UAE. Association for Computational Linguistics.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022. Why do you feel this way? summarizing triggers of emotions in social media posts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 5×10^{-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", "thieves"), 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.

	# topics	# masked topic words
ISEAR	10	100
SSEC	11	110
TALES	10	10
enVENT	8	80
APPReddit	10	100
enISEAR	13	130

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

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 InTopic-gr and CrossTopic-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 predicting 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 λ as it regulates the impact of the GRL. Again, choosing λ is a trade-off between overall classification performance and de-biasing potency. To determine an optimal value for λ, standard emotion (or appraisal classifiers) are trained on each individual corpus for λ 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, λ 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] Suddenness, familiarity, event predictability, pleasantness, unpleasantness, goal relevance, own responsibility, others' responsibility, situational responsibility, anticipation 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.

Corpus	Size	Annotation	Domain	Class. Setting
ISEAR	7666	*joy*, *sadness*, *anger*, *fear*, *disgust*, *shame*, *guilt*	event descr.	single
SSEC	4870	Plutchik	tweets	multi
TALES	10339	Ekman + *no emotion*	fairy tales	single
CROWD-ENVENT	6600	Ekman + *shame*, *pride*, *bored.*, *rel.*, *trust*, *guilt*, *no*	event descr.	single
		21 appraisal dimensions		multi
APPREDDIT	780	*unexp.*, *consist.*, *cert.*, *cntrl.*, *resp.*	reddit posts	multi
ENISEAR	1001	*joy*, *sadness*, *anger*, *fear*, *disgust*, *shame*, *guilt*	event descr.	single
		attent., *cntrl.*, *circum.*, *resp.*, *pleasant.*, *effrt.*, *cert.*		multi

Table 5: Corpus overview. Emotion/appraisal statistics for ENVENT/ENISEAR are reported separately.

reader's and writer's assessment of emotions and appraisals. However, these are not used here.

APPREDDIT. The APPREDDIT corpus (Stranisci et al., 2022) is annotated with appraisal dimensions. It comprises 780 reddit posts, where each posts 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*, *shame*, *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 labelset 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.

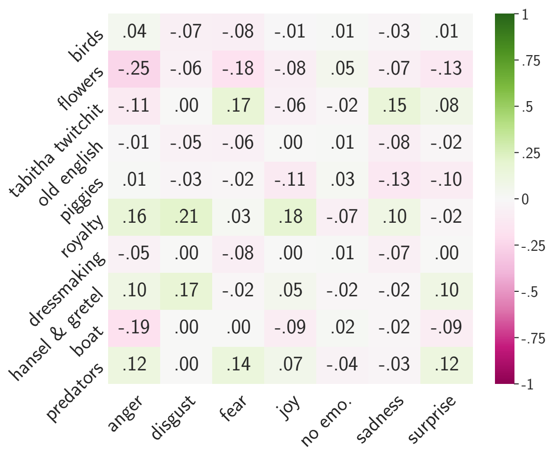

Figure 7: Normalized pointwise mutual information between topics and emotion annotations in TALES.

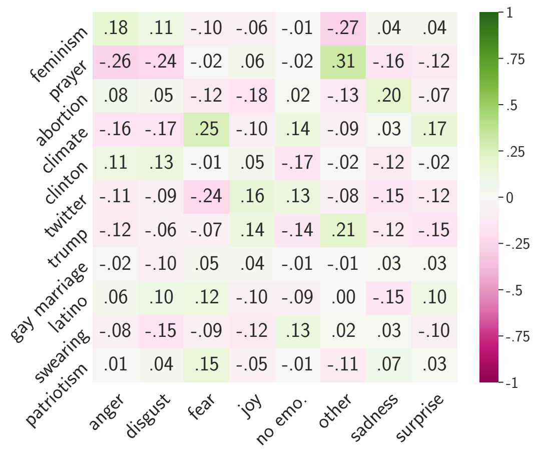

Figure 8: Normalized pointwise mutual information between topics and emotion annotations in SSEC.

Corpus	A	D	F	J	Sa	Sh	Su	No	O
enVENT	550	550	550	550	550	550*	550	550	2,200*
ISEAR	1,096	1,096	1,095	1,094	1,096	2,189*	–	–	–
enISEAR	143	143	143	143	143	286*	–	–	–
SSEC	1388	440	274	815	414	–	177	1552	1077*
TALES	302	40	251	579	340	–	144	8,683	–

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 → joy).

Corpus	Attention	Pleasantness	Suddenness	Self Control	Chance Control	Self Responsibility	Other Responsibility	Predict Consequences	Goal Support	Effort	Other
APPREDDIT	–	–	307	307	–	400	457	748	312	–	–
enVENT	4125	2261	3128	2142	1514	2597	3396	2841	2281	3210	6527*
enISEAR	673	149	–	228	240	377	–	761	–	400	–

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 → joy), or by combining multiple classes into one aggregated).