Towards Procedural Fairness: Uncovering Biases in How a Toxic Language Classifier Uses Sentiment Information

Isar Nejadgholi, Esma Balkır, Kathleen C. Fraser, and Svetlana Kiritchenko
National Research Council Canada
Ottawa, Canada
{Isar.Nejadgholi, Esma.Balkir, Kathleen.Fraser, Svetlana.Kiritchenko}@nrc-cnrc.gc.ca

Abstract

Previous works on the fairness of toxic language classifiers compare the output of models with different identity terms as input features but do not consider the impact of other important concepts present in the context. Here, besides identity terms, we take into account high-level latent features learned by the classifier and investigate the interaction between these features and identity terms. For a multi-class toxic language classifier, we leverage a concept-based explanation framework to calculate the sensitivity of the model to the concept of *sentiment*, which has been used before as a salient feature for toxic language detection. Our results show that although for some classes the classifier has learned the sentiment information as expected, this information is outweighed by the influence of identity terms as input features. This work is a step towards evaluating procedural fairness, where unfair processes lead to unfair outcomes. The produced knowledge can guide debiasing techniques to ensure that important concepts besides identity terms are well-represented in training datasets.

1 Introduction

Previous NLP works have studied the fairness of toxicity detection classifiers by comparing the distributions of prediction scores across different demographic groups as input features (Dixon et al., 2018; Borkan et al., 2019). However, other toxicity-related concepts are often present in the text and affect the differences in score distribution between identity groups. Here, we introduce a framework that uses *concept-based global explanations* to uncover unintended biases for different identity groups, while controlling for a certain toxicity-related concept. To demonstrate the effectiveness of concept-based explanations in uncovering biases, we specifically focus on *sentiment*, although the general methodology can be applied to any other relevant human-defined concept. Negative sentiment is a salient toxicity feature, which has been used in designing feature-based and neural toxicity detection systems (Fortuna and Nunes, 2018; Zhou et al., 2021; Chiril et al., 2022), and highly correlates with toxic language when targeted at demographic groups.

Assessing the differences in score distributions for various demographics is an example of outcome fairness. In fact, most fairness criteria used in machine learning measure outcome fairness, such as accuracy parity (equal accuracy for protected and unprotected groups), equality of opportunity (equal true positive rates), or equalized odds (equal true positive and false positive rates) (Morse et al., 2021). While valuable, outcome fairness metrics are costly to compute as they require large labelled datasets and do not provide any information about the model's decision making processes.

More recently, work has begun to focus on the complementary notion of *process fairness* (also known as *procedural fairness*), or the idea that the decision-making process itself must be fair. Grgic-Hlaca et al. (2016) conducted one of the first studies on process fairness in machine learning, measuring the extent to which people believed it was permissible to use various features as input to a criminal recidivism prediction algorithm. For example, they found that people generally felt that *criminal history* was fair to use as an input feature, but that it was unfair to use *family criminality* as input. Another aspect of process fairness is that the importance given to an attribute in the decision-making process shouldn't be very different for different demographic groups. An example of this is the recent *SFFA vs. Harvard* court case where it was argued that academic and extracurricular achievements of Asian-American applicants are given less weight in the admissions process compared to their White-American counterparts (Arcidiacono et al., 2022). We take a similar view of process fairness and consider a classifier as unfair if it either ignores or over-utilizes a feature for some demographic

groups compared to others.

In the current NLP landscape, one major barrier to assessing process fairness is that predictive models rarely use human-understandable concepts as input features, and so it is increasingly difficult to understand what high-level features[1] are actually being learned and used by the classifier. In this work, we use an interpretability framework of concept-based explanations (Yeh et al., 2022), which enables us to explain a machine learning model's decision-making via conceptual units understandable to humans.

Concept-based explanations have been studied mostly in the context of computer vision, where it is fairly straightforward to define concepts of interest with a set of representative examples. However for textual data, it is much less clear how to define a concept in an effective and intuitive manner, and global explainability methods that operate on high-level abstractions remain under-explored (Danilevsky et al., 2020; Balkır et al., 2022a). Ghorbani et al. (2019) define a concept to be a meaningful, human-defined abstraction, which is expected to be important for the task at hand and which can be specified by a coherent set of examples. Following this definition, we identify *sentiment* as a concept for toxicity classification.[2] To the best of our knowledge, this is one of the first works to apply concept-based explanations to the domain of NLP, and the first one to explore its effectiveness in identifying high-level fairness issues in models that work with textual data.

In this work, we show how to use concept-based explanations to determine whether a trained toxicity classifier uses the information of *sentiment* as an important feature in its predictions. For that we use a multi-class model, described in Section 2, and compare the importance of the concept of sentiment in predicting different subtypes of toxicity. Although intuitively, negative sentiment should be an important signal for toxicity detection, its presence is neither necessary nor sufficient for an utterance to be tagged as toxic. For example, "*Muslims are grieving*" carries a negative sentiment but is not abusive, whereas "*You are so smart for a woman*" is perceived as an insult despite including a positive sentiment word. Also, sentiment might not be a distinguishing feature for some variations or subtypes of toxic language, such as threats or cyberbullying. For all the classes of our multi-class model, we ask, "Has the classifier learned the concept of sentiment as a coherent and important high-level feature associated with this label?", and answer this question with concept-based explanations (Section 5). We then assess how the presence of identity terms impacts the use of sentiment information by the classifier. For that, we control the context for sentiment and ask if the learned sentiment information is used similarly and fairly across identity groups (Section 6). Our code and data is available at https://github.com/IsarNejad/Procedural-Fairness-Sentiment.

Our main contributions are:

- We propose a concept-based explanation framework to determine whether a trained text classifier uses a human-defined concept fairly in its decision making process. To the best of our knowledge, this is the first work that uses concept-based explanations to uncover biases in text classifiers, and the first to formalize concepts with short textual templates.
- To demonstrate the utility of the proposed method, we apply it to a multi-class toxicity classifier and show that when the subject of the sentiment is not specified (e.g., "*They are <SENTIMENT-WORD>*"), the classifier is sensitive to the concept of negative sentiment, for some of the classes.
- Further, we show that when the subject of the sentiment is a specific identity term (e.g., "*<IDENTITY-TERM> are <SENTIMENT-WORD>*"), for some classes, the classifier becomes sensitive to neutral and in some cases even positive sentiment. This demonstrates that the process by which the classifier makes its decision is not the same for all identity groups, and for some groups may even be unfairly associating positive sentiment with toxicity.

2 Multi-Class Toxicity Model

For our experiments, we use an open-source, RoBERTa-based model[3] (Hanu, 2020) trained on the English dataset released as part of a Kaggle competition on identifying and reducing bias in

[1] Here, by "feature" we mean the latent representations of a semantic concept learned by a classifier, as opposed to the low-level input features.

[2] We distinguish between the *concept* of sentiment, as defined by a human through a set of examples, and the *feature* of sentiment, which is implicitly learned by the classifier, although our assumption is that they are closely aligned.

[3] https://huggingface.co/unitary/unbiased-toxic-roberta

toxicity classification of online comments.[4]. The dataset includes public comments from the Civil Comments platform manually annotated for *Toxicity* as well as six toxicity subtypes: *Severe Toxicity*, *Obscene*, *Identity Attack*, *Insult*, *Threat*, and *Sexual Explicit*. The values for each label represent the fraction of the annotators that assigned the label to the comment. There are over 1.8M examples in the training set and around 195K examples in the test set. We exclude the class *Severe Toxicity* from our experiments, since there are only eight training examples with values higher than 0.5 for this class. Further, a subset of the data is annotated for various identity groups mentioned in the text. The most frequently mentioned identity groups include *male*, *female*, *homosexual (gay or lesbian)*, *Christian*, *Jewish*, *Muslim*, *Black*, *white*, *people with psychiatric or mental illness*. The classification model optimizes the competition's official evaluation metric that combines the overall AUC with Bias AUCs for the identity groups (Hanu, 2020). For this, the model's loss function combines the weighted loss functions for two tasks, toxicity prediction and identity prediction. This simple and straight-forward model has been shown to effectively reduce bias on non-toxic sentences that mention identity terms, and results in a competitive score of 93.74 on the test set.

We chose this model for two reasons. First, the model is publicly available and is trained on one of the largest available toxicity dataset, annotated for multiple types of toxicity. An alternative choice for our experiments would be using multiple toxicity classifiers. However, the definitions of subtypes of toxicity are usually ambiguous and similar labels might be used for different subtypes of toxicity across datasets. In the case of our multi-class model, the disparities in using sentiment information can be reliably attributed to differences in subtype definitions. Second, the model is debiased to some extent with regards to outcome fairness metrics. Uncovering biases in such a model highlights the issue that optimizing for outcome fairness does not guarantee the procedural fairness in decision making.

3 Sentiment Lexicon

To formalize sentiment concepts, we employ the NRC Valence, Arousal, and Dominance (NRC-VAD) lexicon (Mohammad, 2018), which provides manually annotated real-valued scores of valence, arousal, and dominance for 20,000 English words. We use the valence scores and convert them into the range from -1 (the most negative) to 1 (the most positive). We automatically select single words from the lexicon that are predominantly used as adjectives in the British National Corpus (BNC)[5] and sort them in decreasing order by their frequency in the BNC. The N most frequent adjectives that can be used to describe humans or groups of humans are manually selected as the sentiment words to define the sentiment concepts. The sentiment range [-1, 1] is divided into five intervals: *very negative* [-1, -0.75], *negative* (-0.75, -0.25), *neutral* [-0.25, 0.25], *positive* (0.25, 0.75), and *very positive* [0.75, 1]. For each interval, $N = 100$ adjectives are selected.[6] These sets of adjectives are then used to populate the sentence templates to define the sentiment concepts as described in Section 5.

4 Concept-Based Explanations

Concept-based explanation is an emerging area in black-box model explainability, aiming to explain neural network models at the abstraction level defined by a human user (Yeh et al., 2022). Most explainability methods provide importance weights for low-level input features such as pixels in images or tokens for text (Sundararajan et al., 2017; Smilkov et al., 2017; Selvaraju et al., 2017; Shrikumar et al., 2017). However, a user might want to evaluate the model's functionality at the level of a concept that is expected to be important for the model's prediction, which can be achieved with concept-based explanations (Koh et al., 2020). Ghorbani et al. (2019) states that a concept needs to satisfy the properties of *meaningfulness*, *coherency* and *importance* for the task at hand. Some examples of concepts in computer vision tasks are the concept of *stripes* for the class of *zebra* (Kim et al., 2018), the concept of *white coat* for the class of *doctor* (Pandey, 2021), and the concept of *nuclei texture* in the detection of tumor tissue in breast lymph node samples (Graziani et al., 2018). In the case of text, Nejadgholi et al. (2022) used concept-based explanations to measure the sensitivity of a

[4] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/

[5] The British National Corpus, version 3 (BNC XML Edition), http://www.natcorp.ox.ac.uk/

[6] The full list of the selected adjectives is available in the Supplemental Material. We also conducted similar experiments with the full NRC-VAD lexicon and obtained similar results.

abusive language classifier to the emerging concept of COVID-related anti-Asian hate speech, and Yeh et al. (2020) explained a text classifier with respect to the concepts identified through topic modeling.

Here, our goal is to explain the prediction of a toxicity classifier at the level of sentiment information learned by the trained model. Since sentiment is not one of the direct input features of the model, feature importance metrics such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) cannot be used to provide its importance. Instead, we consider each level of sentiment as a concept and calculate the importance of these concepts for the model's predictions with Testing Concept Activation Vectors (TCAV) algorithm. In the following section, we explain the TCAV algorithm in detail.

4.1 Testing Concept Activation Vectors

Testing Concept Activation Vectors (TCAV) is an algorithm from the family of concept-based explainability methods, which measures the importance of a human-defined concept for model's predictions (Kim et al., 2018). In TCAV, each concept is defined with a set of examples and represented as Concept Activation Vectors (CAVs), in the activation layer of the trained model. TCAV formalizes the importance of a concept as the fraction of input examples for which the prediction scores of the model increase if the input representation is infinitesimally moved towards the concept representation. The prediction increase is measured by calculating the directional derivatives of the prediction layer to CAVs. To calculate the statistical significance of a concept, multiple subsets of concept examples are used to form multiple CAVs, and a TCAV score is calculated for each CAV. A concept is considered to be important for a class if the distribution of its TCAV scores is significantly different from the TCAV scores of a random concept defined by random examples.

Here, we explain how the TCAV procedure measures the importance of a concept for a class of a RoBERTa-base classifier, in more detail. Similar to Nejadgholi et al. (2022), we define each concept C with N_C concept examples, and map them to RoBERTa representations of the [CLS] token $r_C^j, j = 1, ..., N_C$. Then, P number of Concept Activation Vectors (CAVs), v_C^p, are generated by averaging the RoBERTa representations of N_v randomly chosen concept examples, to represent C in the activation space:

$$v_C^p = \frac{1}{N_v} \sum_{j=1}^{N_v} r_C^j \quad p = 1, .., P \tag{1}$$

where $N_v < N_C$. With f_{emb}, which maps an input text x to its RoBERTa representation r_x, the *conceptual sensitivity* of a class to the v_C^p, at input x can be computed as the directional derivative $S_{C,p}(x)$:

$$S_{C,p}(x) = \lim_{\epsilon \to 0} \frac{h(f_{emb}(x) + \epsilon v_C^p) - h(f_{emb}(x))}{\epsilon}$$
$$= \bigtriangledown h(f_{emb}(x)).v_C^p \tag{2}$$

where h is the function that maps the RoBERTa representation to the logit value of the class of interest. For a set of input examples, X, we calculate the TCAV score as the fraction of inputs for which small changes in the direction of C increase the logit:

$$TCAV_{C,p} = \frac{|x \in X : S_{C,p}(x) > 0|}{|X|} \tag{3}$$

When calculated for all CAVs, Equation 3 results in a distribution of scores for the concept C. The mean and standard deviation of this distribution determines the overall sensitivity of the classifier to the concept C for the class of interest.

Intuitively, the derivatives in Equation 2 indicate whether a label's likelihood increases when a small vector in the direction of the concept's representation is added to a random context. For example, the predicted probability of the class *Toxic* for sentence *"I saw these people."* is 0.01. The comment *"I saw these people. They are terrible."* is labeled as toxic with the probability of 0.56, but the statement *"I saw these people. They are wonderful."* receives the toxicity probability of 0.01. If this observation holds systematically across many negative and positive sentiment words, the classifier has learned negative sentiment as an important feature of the toxicity class, but the positive sentiment does not contribute to the toxicity estimation.

In contrast to the previous concept-based explanation works in NLP, which either require annotated data (Nejadgholi et al., 2022), or are limited to the topics extracted by the topic modeling procedure (Yeh et al., 2020), we define the sentiment concepts with a set of minimal templates, that are easy to generate and minimize extra contextual information. Using concept examples, described in Sections 5 and 6, TCAV first encodes the information of sentiment in the RoBERTa embedding

Class label	Sentiment level concepts					Control concepts	
	Very negative	Negative	Neutral	Positive	Very positive	Explicit	Non-coherent
Toxicity	**0.87 (0.04)**	**0.47 (0.26)**	0 (0)	0 (0)	0 (0)	**0.88 (0.02)**	0 (0)
Obscene	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	**0.75 (0.1)**	0 (0)
Identity Attack	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.01 (0.03)	0 (0)
Insult	**0.92 (0.02)**	**0.77 (0.14)**	0 (0)	0 (0)	0 (0)	**0.89 (0.02)**	0 (0)
Threat	0.01 (0.03)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Sexual Explicit	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	**0.70 (0.18)**	0 (0)

Table 1: Means and standard deviations of TCAV score distributions for the six types of toxicity with respect to five sentiment categories and two control concepts. Scores statistically significantly different from random are in bold.

space. Then, it populates the directional derivatives of the prediction layer with respect to these vectors. If the derivatives are positive for a significant number of the sentiment concept representations, a high average TCAV score is obtained, i.e, sentiment is learned as a coherent and important feature for the label of interest.

In our implementation, for each concept, 100 CAVs are generated, each of which is the average representation of 50 randomly selected concept examples. For 1000 random input texts (random tweets collected with stop words), the TCAV scores of each CAV are calculated. Average and standard deviation of TCAV scores are reported to quantify the importance of the concept for this class.

5 Classifier Sensitivity to Sentiment

In this section, we analyze the sensitivity of the classifier described in Section 2 and identify the classes for which the classifier is sensitive to sentiment as a coherent and important feature. For the concepts of sentiment level, we create concept examples and use the TCAV technique to test the importance of the concepts for each class.

To create concept examples for each level of sentiment described in Section 3, we use the template "*They are <SENTIMENT-WORD>.*". This simple template ensures minimal extra semantic information other than sentiment level and avoids the problem of encoding unwanted biases in the concept itself, as described by Tong and Kagal (2020). It is important to note that such minimal templates cannot be labeled for toxicity without more context. Consider a sentence such as "*They are terrible.*". This sentence expresses a negative sentiment but lacks any other significant information. Only with more context can we say whether this sentence is toxic or not. The statement "*These people are immigrants. They are terrible.*" is toxic, while the comment "*I don't like these computers. They are terrible.*" is non-toxic, and "*I don't like these singers. They are terrible.*" can be toxic depending on the specific definition of toxicity in a use case. Therefore, the fairness analysis methods that rely on labels cannot be used to study the impact of these templates on the model's predictions.

In addition to five levels of sentiment, we use two control concepts with predictable sensitivities: 1) A non-coherent concept, defined by a set of random tweets collected with stop words, for which we expect low average TCAV scores for all labels; 2) The concept of *explicit offence* defined by inserting a profane word[7] in the template "*They are <PROFANE-WORD>*", for which we expect high sensitivity from at least some of the labels. As the creators of the toxicity model mention, this classifier shows high sensitivity to profanity because of the over-representation of these words in its training data.[8] Table 1 shows the average and standard deviation of TCAV scores calculated for the seven concepts described above.

We observe that the TCAV scores for the control concepts are as expected—zero sensitivity for a non-coherent, random concept and high sensitivity to the concept of explicit offence for the labels *Toxicity*, *Obscene*, *Insult* and *Sexual Explicit*. For the sentiment concepts, we observe that the classifier is sensitive to *Very Negative* and *Negative* sentiment for the labels *Toxicity* and *Insult*. [9] We also observe that the classifier is not sensitive to the *Neutral*, *Positive* and *Very Positive* sentiment concepts for any of the labels, which rules out the sensitivity of the classifier to the specific sentence structure of the templates.

Literature suggests that a high TCAV score indicates: 1) the concept is learned by the classifier as

[7] We use the words from https://github.com/chucknorris-io/swear-words/blob/master/en

[8] https://github.com/unitaryai/detoxify

[9] Intuitively, the classes *Obscene*, *Identity Attack*, *Threat* and *Sexual Explicit* rely on features other than negative sentiment, i.e., profanity, identity terms, violence or intention of harming, and lewdness, respectively.

Class label	Sentiment level concepts					Control concepts	
	Very negative	Negative	Neutral	Positive	Very positive	Explicit	Non-coherent
Toxicity	**0.22**	**0.12**	0.01	0	-0.01	**0.27**	0.03
Obscene	0.01	0	0	0	0	**0.10**	0
Identity Attack	0.01	0	0	0	0	0.03	0
Insult	**0.17**	**0.10**	0.02	0	0	**0.16**	0
Threat	0	0	0	0	0	0	0
Sexual Explicit	0	0	0	0	0	**0.09**	0

Table 2: Average increase in probabilities when concept templates are added to random texts. Cells in equivalent positions to Table 1 are in bold.

Class label	Increase in Probability		TCAV scores	
	non-coherent	coherent	non-coherent	coherent
Toxicity	0.11	0.12	0.01 (0.07)	**0.47 (0.26)**
Insult	0.09	0.10	0.09 (0.19)	**0.77 (0.14)**

Table 3: Average increase in probability and mean and standard deviation of TCAV scores for the non-coherent concept (Very negative or Very positive sentiment) and the coherent concept (Negative sentiment).

a coherent feature, and 2) that feature is important for the classifiers' predictions (Kim et al., 2018). We evaluate the TCAV scores shown in Table 1, in terms of the *importance* and *coherency* of a concept. We first confirm that the *importance* of a concept can be interpreted as the *increase in the predicted probability due to the addition of a concept to input sentences*. Then, we show that *increase in probability* is not an equivalent metric to *TCAV score*, since increase in probability can be due to the addition of a non-coherent concept to input sentences.

High average TCAV scores indicate a significant increase of prediction probabilities when the concept is added to random contexts. We append the concept examples to random tweets and measure the prediction probabilities before and after the addition of the concept examples. The average increase of probabilities for all concepts and labels is shown in Table 2. We observe that in all cases where the average TCAV scores are high (i.e., significantly different from the control random concept) in Table 1, the probability increase is notable in Table 2. For example, for the *Toxic* label, the addition of sentences with *Very Negative* and *Negative* sentiment on average increases the prediction probability by 0.22 and 0.12, whereas the addition of *Neutral*, *Positive* and *Very Positive* sentiments increases the prediction probability by 0.01 or less. This is in line with our observation from Table 1 that the classifier is sensitive to *Negative* and *Very Negative* sentiments for the label *Toxic*.

TCAV scores differentiate between coherent and non-coherent concepts whereas the probability increase does not. To test this hypothesis, we create a non-coherent concept by combining the *Very Negative* and the *Very Positive* sentiment examples, and compare the average increase in probability and the TCAV score for this concept with those for the *Negative* sentiment concept. The comparison is demonstrated in Table 3. Although the increase in probability is similar for the coherent and non-coherent concepts, the TCAV score indicates that the classifier has only learned the coherent concept as an important feature.

6 Sensitivity to Sentiment Towards an Identity Group

In the previous section, we demonstrated how the TCAV framework can be used to assess whether a human-defined concept is learned by a classifier as an important feature. With that we showed that for some labels our model is sensitive to the presence of *Very Negative* and *Negative* sentiments in broader contexts. Here, we turn to the concept of *"associating a sentiment with an identity group"*[10] and ask if similar levels of sensitivity to sentiment are observed in the presence of certain demographic terms as input features. For creating the concept examples, we use the template "<SUBJECTS> are <SENTIMENT-WORD>", where <SUBJECTS> are the protected identity terms used in HateCheck (Röttger et al., 2021): *Women*, *Gay people*, *Trans people*, *Muslims*, *Immigrants*, *Black people*, and *Disabled people*. We also add

[10]Note that this concept is composed of more basic concepts, similar to the concept of *white coat* used in (Pandey, 2021). Still, it satisfies the three criteria of meaningfulness, coherency and importance as stated by (Ghorbani et al., 2019) and can be considered as a relevant concept for toxicity.

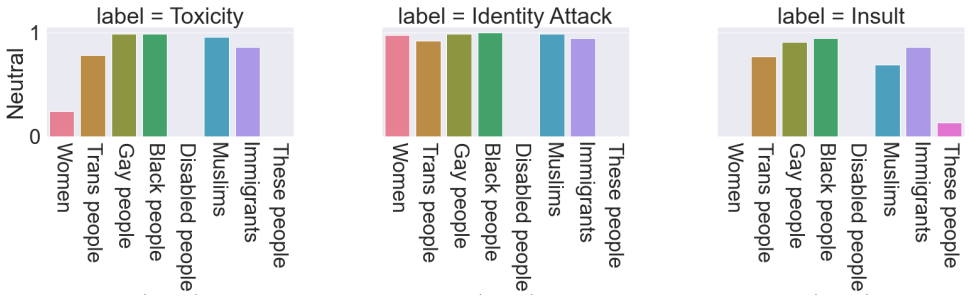

Figure 1: Sensitivity to identity terms in neutral contexts, for which a low sensitivity is expected.

the terms *These people* and *These things*, to assess the sensitivity of the model to the concepts of *"associating a sentiment with people in general"* and *"associating a sentiment with objects"* as two baselines. As expected, we observe that the classifier is not sensitive to any level of sentiment when associated with objects. We discuss some of the most salient results below. (The full results are presented in Appendix in Table A.1.)

We first assess the influence of identity terms by analysing the classifiers' sensitivities to the neutral sentiment towards the identity groups. Figure 1 visualizes the column of Table A.1 related to the *Neutral* sentiment. From the findings in Table 1, as well as human intuition, the association of identity terms with neutral sentiment should not increase the probability of the classifier predicting a toxic label. However, we observe high sensitivities for the labels *Toxicity* and *Insult* and all identity groups, except for *Women* and *Disabled people*. The classifier is also sensitive to the *Neutral* sentiment associated with *Women* for the label *Identity Attack*. We conclude that in neutral contexts the classifier is more sensitive to some identity terms than others.

Figure 2 visualizes the results of Table A.1 from a different perspective. In this figure, we assess the sensitivity of the classifier to different levels of sentiment across the identity groups. For the relevant classes, we expect to see that the classifier is sensitive to negative sentiment but not sensitive to positive sentiment; i.e., the average TCAV score should be high for negative sentiment and low for positive sentiment. Consistent with results in Table 1, we observe this expected pattern for the subject *These people* and the two classes *Insult* and *Toxicity*. Taking this as our baseline, we expect to see similar patterns across all the identity groups for an unbiased classifier. However, we observe that the pattern only holds for certain classes, and certain identity groups. Notably, the classifier loses its sensitivity to the *Negative* sentiment for the classes *Insult* and *Toxicity* when the mentioned identity group is *Disabled people*. In addition to classes *Insult* and *Toxicity*, in the presence of an identity term, the classifier becomes sensitive to *Negative* and *Very Negative* sentiment for the class *Identity Attack*. This is expected given the class definition, but again the pattern does not hold for the identity term *Disabled people*.

Another anomalous result with respect to the *Negative* sentiment is that the classifier is sensitive to *Negative* and *Very Negative* sentiment for the class *Sexual Explicit*, but only in the context of the identity group *Gay people*. Additionally, for the label of *Threat* the classifier shows some sensitivity to the *Very Negative* and *Negative* sentiment, but only when used with the identity term *Muslims*.

We then turn to assessing the sensitivities to positive sentiment. In general, the expected pattern of sensitivities is only observed for the labels *Toxicity* and *Insult* and the identity term *Women*. For other cases, as the level of sentiment changes from *Very Negative* and *Negative* to *Positive* and *Very Positive*, the sensitivity of the classifier does not decrease as expected and remains high, specifically for identity groups *Gay people*, *Black people* and *Muslims*. This observation indicates that a sentence such as "*Black people are wonderful*" in a conversation will increase the probability of that text being labeled as *Toxic*, *Identity Attack* and *Insult*. Also, for *Sexual Explicit* the classifier is sensitive to the mention of *Gay people* for all levels of sentiment. We conclude that the classifier is oversensitive to the presence of these identity terms regardless of the level of sentiment, and even highly positive sentiment cannot cancel out the impact of the identity terms. One exception is the term *Disabled people*, for which the model is under-sensitive, i.e, not sensitive to even *Negative* and *Very Negative* sentiments associated with this group.

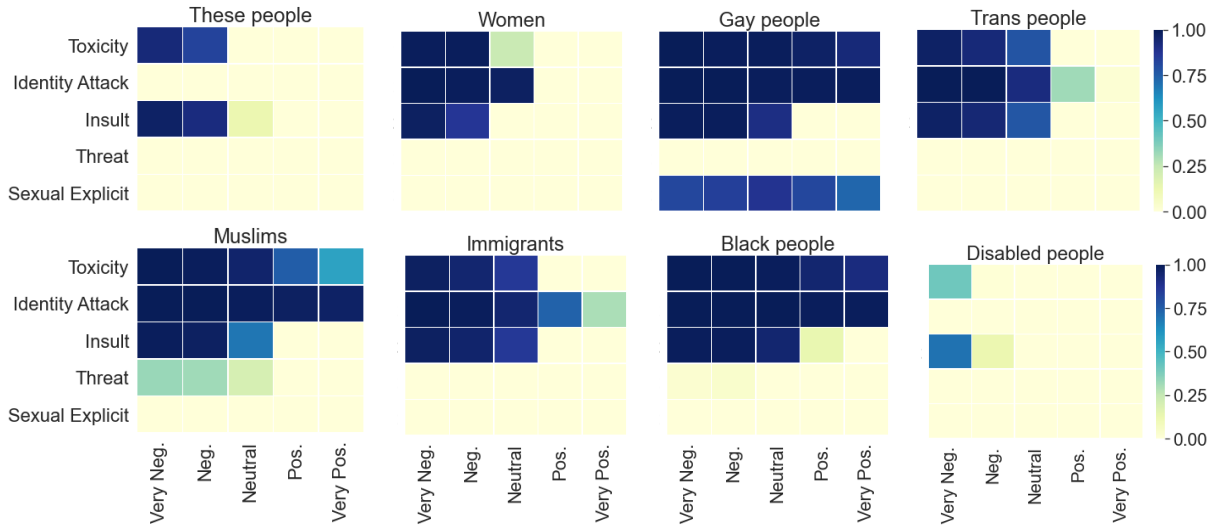

Figure 2: Sensitivity to various levels of sentiment for all demographic groups.

7 Discussion

Our results demonstrate that while a multi-class toxicity classifier generally shows high sensitivities to negative sentiment for certain classes, and zero sensitivities to neutral or positive sentiment, the picture changes when the sentiment is applied to certain marginalized identity groups. Then, counter-intuitively, even positive sentiment can increase the probability of a toxicity label. This suggests an over-reliance on the identity group term.

Previous work in computer vision has underscored the difficulty in finding unbiased examples with which to define concepts, e.g., when searching for images of men or women to examine gender bias, the examples invariably also contain information about age, race, and so on. Here, in the context of NLP, we propose a generalizable solution to that problem, by generating examples rather than collecting them, and carefully controlling the variable of interest (here, sentiment, although the method could extend to other features). For this we use existing lexicons, without the need to label the examples for the various toxicity classes, as would be required for an analysis of outcome fairness.

This knowledge of how the model uses the sentiment information can guide debiasing techniques. For example, a data augmentation approach can ensure important features are present in the training dataset. In the case of our model, a data balancing procedure should collect and label positive and very positive sentiments associated with gay people, Black people and Muslims, as well as very negative sentiments associated with disabled people and add them to the training dataset.

It is important to note that evaluating models for the sensitivity to human-defined concepts is a tool to reveal flaws of a trained model, where prior knowledge about expected sensitivities is available. Similar to previous test suits such as HateCheck (Röttger et al., 2021), our method should not be considered as a standalone evaluation of models. Moreover, observing the expected sensitivities does not guarantee the fairness of the model. Only where unexpected sensitivity patterns are observed, the biases can be detected and mitigated accordingly.

Our method has limitations. We carry our analysis for one grammatical construction that expresses the concept of *associating sentiment to identity groups*. Future work is needed to assess the generalizability of our results to other expressions of sentiment. Moreover, TCAV requires access to at least some model layers and cannot be applied when the model itself is unavailable.

8 Related work

Identifying and mitigating unintended biases in NLP systems to ensure fair treatment of various demographic groups has been focus of intensive research in the past decade (Blodgett et al., 2020; Shah et al., 2020). Various metrics to quantify biases in system outputs have been proposed, including group fairness metrics and individual fairness metrics (Castelnovo et al., 2022; Czarnowska et al., 2021). However, to apply such metrics, the datasets need to be annotated with demographic attributes, which is costly and sometimes infeasible to do (e.g., the demographics of the authors of social media

posts are often unknown). Alternatively, the bias metrics are applied on synthetic data automatically generated using simple templates (Kiritchenko and Mohammad, 2018; Borkan et al., 2019). In both cases, the test data are limited, and the evaluation is restricted to a set of pre-defined contexts.

Explainability techniques (XAI) can potentially help in discovering and quantifying biases. Much work on XAI has been motivated by the need to assist in bias detection and mitigation (Doshi-Velez and Kim, 2017; Das and Rad, 2020). However, only a handful of NLP studies have actually employed explainability methods for bias detection and to a limited extent (Prabhakaran et al., 2019; Kennedy et al., 2020; Aksenov et al., 2021; Balkır et al., 2022b). Balkır et al. (2022a) surveyed the works at the intersection of fairness and XAI in NLP and discussed conceptual and practical challenges in applying current explainability approaches for debiasing NLP models. Multiple outlined issues stem from the fact that most current XAI methods employed in NLP provide explanations on a local level through post-hoc processing, and it is still an open question how to generalize these local explanations to reveal systematic model biases. The TCAV framework used in this paper produces global explanations and can therefore uncover unfair processes in the model's decision making.

Probing classifiers are well-known interpretability tools used to examine the encoded information in the representation layers of NLP models (Conneau et al., 2018). Probes are trained independently from the original model to predict an externally defined property (e.g., linguistic properties such as part of speech) from the model's representations. Despite being widely used, several studies revealed that probes are not well-controlled, and caution should be taken when drawing behavioural conclusions about the original model from the performance of probing classifiers (Belinkov, 2022). Also, probes can only assess whether the information about the property of interest is encoded in the representations (e.g., (Tenney et al., 2019; Rogers et al., 2020)) but do not provide evidence about how the model uses this information. To that end, extensions of probing classifiers were proposed, which assess the effect of removing the property's information with counterfactual interventions to provide causal explanations and mitigate biases in NLP classifiers (Ravfogel et al., 2020; Elazar et al., 2021). However, several concerns are raised about the effectiveness and the unintended consequences of removing attributes (Kumar et al., 2022). While causality-driven probing methods assess the necessity of the property for the classifier's decision, TCAV determines whether the model uses the encoded information as an important signal for a particular class. Also, TCAV allows us to quantify the relative importance of different properties encoded in the representation, which is not feasible with probing classifiers.

The TCAV framework has been developed and mostly applied in image classification. In the original paper, Kim et al. (2018) showed how gender and racial biases can be discovered with TCAV in image classifiers. Wei et al. (2021) extended the method to regression problems, and applied it to detect gender and first language biases in automatic spoken language assessment. Tong and Kagal (2020) studied the effectiveness of TCAV in discovering gender biases in image classification and discussed the difficulties in obtaining quality examples to represent a concept while not introducing new sources of bias (e.g., introducing a racial bias when selecting gendered examples). Adhikari (2021) used TCAV to measure gender bias when classifying faces as young or old, and discussed the difficulty of defining 'disentangled' concepts that only encode the concept of interest. To the best of our knowledge, our work is the first in applying TCAV to discover biases in text classifiers.

9 Conclusion

Building on previous studies that measured group fairness in toxic language detection, this work is a step toward a more systematic and fine-grained analysis of procedural fairness in neural model's predictions. We use a global explainability metric to uncover the disparities in how the classifier learns to associate identity terms with domain-relevant concepts, e.g. sentiment. Future work will focus on extending the analysis to other concepts known to be important to toxic language detection (profanity, threats of violence, dehumanizing or othering language, and so on) as well as additional classifiers, domains, and types of bias.

Ethical Statement

The presented framework aims to identify fairness issues in text classifiers when identity terms are mentioned in the text. As stated above, such evaluation cannot attest for the absence of any biases,

but can indicate potential areas of concern. This framework is a complementary approach to other methods of bias detection that are based on the notion of outcome fairness (e.g., using fairness metrics on held-out test sets annotated for mentions of demographics or on specifically designed test suits, such as HateCheck). The proposed method cannot be applied to assessing fairness on texts *written by* different demographic groups.

The method requires the identity groups of interest to be specified in advance. In the current study, we have included several protected groups, but the list is by no means exhaustive. More protected groups should be included in the future. Additionally, it is known that the label used to refer to a social group can itself communicate bias (consider, for example, the difference between *immigrants* versus *migrants* versus *expats*) (Beukeboom and Burgers, 2019). We have not analyzed the effect of this form of bias on the explanations here. Furthermore, other, legally non-protected groups (e.g., based on physical appearances, education, etc.), should also be considered as we strive towards inclusive and safe online spaces.

As most AI technology, this approach can be used adversely to exploit the system's vulnerabilities and produce toxic texts that would be undetectable by the studied classifier. Specifically, for methods that require access to the model's inner layers, care should be taken so that only trusted parties could gain such access. The obtained knowledge should only be used for model transparency purposes, and the security concerns should be adequately addressed.

Regarding environmental concerns, contemporary NLP systems based on pre-trained large language models, such as RoBERTa, require significant computational resources to train and even fine-tune. Larger training datasets, such as the one used in this study with almost 2M training examples, used for fine-tuning, usually result in a better classification performance, but also an even higher computational cost. To lower the cost of this study and its negative impact on the environment, we chose to use an existing, publicly available classification model.

References

Rittika Adhikari. 2021. Fair-doctor: Detecting and mitigating unfairness in neural networks. Master's thesis, University of Illinois Urbana-Champaign.

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno Schneider, and Georg Rehm. 2021. Fine-grained classification of political bias in German news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131.

Peter Arcidiacono, Josh Kinsler, and Tyler Ransom. 2022. Asian american discrimination in harvard admissions. *European Economic Review*, 144:104079.

Esma Balkır, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2022a. Challenges in applying explainability methods to improve the fairness of NLP models. In *Proceedings of the Second Workshop on Trustworthy Natural Language Processing (TrustNLP @ NAACL)*, Seattle, WA, USA.

Esma Balkır, Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022b. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2672–2686, Seattle, United States. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, 7:1–37.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 491–500.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, 14(1):322–352.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459.

Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32.

Mara Graziani, Vincent Andrearczyk, and Henning Müller. 2018. Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 124–132. Springer.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *Proceedings of the NIPS Symposium on Machine Learning and the Law*, volume 1, page 2. Barcelona, Spain.

Laura Hanu. 2020. How well can we detoxify comments online? Unitary, accessed on 15 June, 2022.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the International Conference on Machine Learning*, pages 2668–2677. PMLR.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning*, pages 5338–5348. PMLR.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. Probing classifiers are unreliable for concept removal and detection. *arXiv preprint arXiv:2207.04153*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Lily Morse, Mike Horia M Teodorescu, Yazeed Awwad, and Gerald C Kane. 2021. Do the ends justify the means? variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics*, pages 1–13.

Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Improving generalizability in implicitly abusive language detection with concept activation vectors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5517–5529, Dublin, Ireland. Association for Computational Linguistics.

Parul Pandey. 2021. Tcav: Interpretability beyond feature attribution. Breaking the Jargons, accessed on 13 June, 2022.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection 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 41–58, Online. Association for Computational Linguistics.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

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.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning*, pages 3145–3153.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, pages 3319–3328.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Schrasing Tong and Lalana Kagal. 2020. Investigating bias in image classification using model explanations. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI 2020)*.

Xizi Wei, Mark JF Gales, and Kate M Knill. 2021. Analysing bias in spoken language assessment using concept activation vectors. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7753–7757. IEEE.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565.

Chih-Kuan Yeh, Been Kim, and Pradeep Ravikumar. 2022. Human-centered concept explanations for neural networks. In P. Hitzler and M. K. Sarker, editors, *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342, page 2. IOS Press.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021. Hate speech detection based on sentiment knowledge sharing. 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 7158–7166, Online. Association for Computational Linguistics.

A Sensitivities to Sentiment in Presence of Identity Terms

The full results of the experiments described in Section 6 are presented in Table A.1.

Target	Class label	Very negative	Negative	Neutral	Positive	Very positive
Women	*Toxicity*	**0.99(0.00)**	**0.99(0.00)**	0.24(0.22)	0(0)	0(0)
	Obscene	0(0)	0(0)	0(0)	0(0)	0(0)
	Identity Attack	**1.00(0)**	**0.99(0.00)**	**0.98(0.02)**	0(0.01)	0(0)
	Insult	**0.98(0.01)**	**0.87(0.11)**	0(0)	0(0)	0(0)
	Threat	0(0)	0(0)	0(0)	0(0)	0(0)
	Sexual Explicit	0(0)	0(0)	0(0)	0(0)	0(0)
Trans people	*Toxicity*	**0.97(0.01)**	**0.93(0.01)**	**0.78(0.04)**	0(0)	0(0)
	Obscene	0(0)	0(0)	0(0)	0(0)	0(0)
	Identity Attack	**1.00(0)**	**1.00(0)**	**0.92(0.01)**	**0.32(0.20)**	0.02(0.05)
	Insult	**0.97(0.007)**	**0.94(0.01)**	**0.77(0.07)**	0(0)	0(0)
	Threat	0(0)	0(0)	0(0)	0(0)	0(0)
	Sexual Explicit	0(0)	0(0)	0(0)	0(0)	0(0)
Gay people	*Toxicity*	**1.00(0)**	**0.99(0.00)**	**0.99(0.00)**	**0.97(0.00)**	**0.93(0.01)**
	Obscene	0.25(0.12)	0.01(0.02)	0(0)	0(0)	0(0)
	Identity Attack	**1.00(0)**	**1.00(0)**	**0.99(0.00)**	**0.99(0.00)**	**0.99(0.00)**
	Insult	**0.99(0.001)**	**0.99(0.003)**	**0.91(0.04)**	0(0)	0(0)
	Threat	0(0)	0(0)	0(0)	0(0)	0(0)
	Sexual Explicit	**0.82(0.02)**	**0.84(0.01)**	**0.88(0.01)**	**0.82(0.02)**	**0.73(0.02)**
Black people	*Toxicity*	**1.00(0)**	**1.00(0)**	**0.99(0.00)**	**0.95(0.00)**	**0.92(0.01)**
	Obscene	0.05(0.06)	0.00(0.00)	0(0)	0(0)	0(0)
	Identity Attack	**1.00(0)**	**1.00(0)**	**1.00(0)**	**0.99(0.00)**	**0.99(0.00)**
	Insult	**0.99(0.001)**	**0.99(0.002)**	**0.95(0.01)**	0.14(0.12)	0(0)
	Threat	0.03(0.02)	0.04(0.02)	0(0)	0(0)	0(0)
	Sexual Explicit	0(0)	0(0)	0(0)	0(0)	0(0)
Disabled people	*Toxicity*	**0.41(0.2)**	0.01(0.06)	0(0)	0(0)	0(0)
	Obscene	0(0)	0(0)	0(0)	0(0)	0(0)
	Identity Attack	0(0)	0(0)	0(0)	0(0)	0(0)
	Insult	**0.70(0.2)**	0.13(0.21)	0(0)	0(0)	0(0)
	Threat	0(0)	0(0)	0(0)	0(0)	0(0)
	Sexual Explicit	0(0)	0(0)	0(0)	0(0)	0(0)
Muslims	*Toxicity*	**1.00(0)**	**0.99(0.00)**	**0.96(0.01)**	**0.75(0.04)**	**0.57(0.07)**
	Obscene	0(0.02)	0(0)	0(0)	0(0)	0(0)
	Identity Attack	**1.00(0)**	**1.00(0)**	**0.99(0.00)**	**0.98(0.00)**	**0.97(0.00)**
	Insult	**0.99(0.00)**	**0.98(0.007)**	**0.69(0.15)**	0(0)	0(0)
	Threat	**0.33(0.07)**	**0.32(0.07)**	**0.20(0.06)**	0(0)	0(0)
	Sexual Explicit	0(0)	0(0)	0(0)	0(0)	0(0)
Immigrants	*Toxicity*	**0.98(0)**	**0.95(0.01)**	**0.86(0.03)**	0(0)	0(0)
	Obscene	0(0)	0(0)	0(0)	0(0)	0(0)
	Identity Attack	**1.00(0)**	**0.99(0)**	**0.95(0.01)**	**0.74(0.11)**	**0.30(0.23)**
	Insult	**0.98(0.005)**	**0.96(0.01)**	**0.86(0.03)**	0(0)	0(0)
	Threat	0(0)	0(0)	0(0)	0(0)	0(0)
	Sexual Explicit	0(0)	0(0)	0(0)	0(0)	0(0)
These people	*Toxicity*	**0.93(0.02)**	**0.83(0.07)**	**0(0.03)**	0(0)	0(0)
	Obscene	0(0)	0(0)	0(0)	0(0)	0(0)
	Identity Attack	0(0)	0(0)	0(0)	0(0)	0(0)
	Insult	**0.97(0.01)**	**0.92(0.02)**	0.13(0.21)	0(0)	0(0)
	Threat	0(0)	0(0)	0(0)	0(0)	0(0)
	Sexual Explicit	0(0)	0(0)	0(0)	0(0)	0(0)
These things	*Toxicity*	0(0)	0(0)	0(0)	0(0)	0(0)
	Obscene	0(0)	0(0)	0(0)	0(0)	0(0)
	Identity Attack	0(0)	0(0)	0(0)	0(0)	0(0)
	Insult	0(0.03)	0(0)	0(0)	0(0)	0(0)
	Threat	0(0)	0(0)	0(0)	0(0)	0(0)
	Sexual Explicit	0(0)	0(0)	0(0)	0(0)	0(0)

Table A.1: Average and standard deviation of TCAV scores for all the labels and different levels of sentiment ranging from Very Negative to Very Positive for the template **"<SUBJECTS> are <SENTIMENT-WORD>"**. All the sensitivities that are significantly different from random are in bold.