Counterfactual Replacement Analysis for Interpretation of Blackbox Sexism Classification Models

Anders Knospe

Dartmouth College, anders.h.knospe.23@dartmouth.edu

Follow this and additional works at: https://digitalcommons.dartmouth.edu/cs_senior_theses

Part of the Computer Sciences Commons

Recommended Citation

https://digitalcommons.dartmouth.edu/cs_senior_theses/19

This Thesis (Undergraduate) is brought to you for free and open access by the Computer Science at Dartmouth Digital Commons. It has been accepted for inclusion in Computer Science Senior Theses by an authorized administrator of Dartmouth Digital Commons. For more information, please contact dartmouthdigitalcommons@groups.dartmouth.edu.
Counterfactual Replacement Analysis for Interpretation of Blackbox Sexism Classification Models

Anders Knospe
Department of Computer Science, Dartmouth College
anders.h.knospe.23@dartmouth.edu

Abstract
This paper describes the AKD team’s system designed for SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS). We implement a simple fine-tuned GPT-3 model, ranking 26 on the leaderboard for task A. We also discuss different approaches to interpretability in the context of critiquing the EDOS task’s sub-category oriented approach. Finally, we propose counterfactual replacement analysis, a novel prototype technique for approaching explainability.

Content warning: This paper examines examples of abject sexism, some of which contains strong language.

The Explainable Detection of Online Sexism (EDOS) task (Kirk et al., 2023) addresses the twofold task of classifying both whether or not a text post is sexist and providing insight into why the text post is sexist. The task involves three classification sub-tasks of varying degrees of granularity. Task A involves binary detection of sexism, task B classification into four “categories” of sexism, and task C classification of eleven “fine-grained vectors” of sexism. These tasks are hierarchical; non-sexist posts are not considered for tasks B and C, and each vector in task C fits into a singular category in task B. The stated rationale for classifying sub-categories of sexism is that doing so will contribute to both accuracy and explainability. Similar sexism detection tasks involve binary sexism detection (Shushkevich and Cardiff, 2019), classifying texts as sexist and/or aggressive (Safi Samghabadi et al., 2020), and classifying first-person accounts of sexist actions (Abburi et al., 2020). Other classification tasks with multiple subcategories of sexism use different subcategory taxonomies (Liu et al., 2018; Sharifirad and Jacovi, 2019).

This paper proceeds in three related directions. First, we implement Bi-LSTM, BERT, and GPT-3 approaches to the problem of binary classification (task A), and find that an unoptimized fine-tuned GPT-3 model yields surprisingly impressive results. We then fine-tune the same GPT-3 architecture on the subcategory classification tasks (B and C) and obtain acceptable performance. See section 2.

Second, we discuss issues surrounding interpreting black-box language models, with a focus on the task introduced by EDOS. In particular, we argue that EDOS’ focus on using subcategories for explainability has serious theoretical drawbacks. See section 3.

Third, we provide a prototype for an alternative approach to interpreting black-box binary classification language models. This techniques examines a model’s behavior on counterfactual passages with slight semantic changes to get at how a model “understands” a complex concept (in this case, sexism). See section 4.

1 The EDOS Task
The EDOS involves three sub-tasks: binary classification as sexist or non-sexist, classification into one of four categories, and classification into one of eleven vectors. The texts in question are short posts sourced from Reddit and Gab’s forum websites. Example posts include:

“wow that’s fucking bullshit. i can’t stand when a woman says "girls can’t rape guys wtf" bonus points if they claim the guy always wants it.”

“men without a women can think and be succesfull”

The task hosts label the texts as sexist or non-sexist. If they are sexist, their two levels of subcategory (e.g., sexist; derogation; descriptive attacks or sexist; animosity; casual use of gendered slurs).

2 Our Approach
In this section, we examine the performance of several straightforward models on EDOS task A.
The model architectures we discuss are trained and evaluated on an 80-20 train-test-split of the EDOS training dataset, consisting of 14,000 short text posts sourced from Reddit and Gab. In all cases, pre-processing was limited to the lower casing.

2.1 Task A

2.1.1 Models

Bi-LSTM + Attention Bi-LSTM with attention model (Graves and Schmidhuber, 2005; Lai et al., 2015) is well-suited for handling sequential data like text. We implemented this model with GloVe embeddings (Pennington et al., 2014).

BERT: BERT models (Devlin et al., 2018) based on the transformer architecture are another commonly used technique. In our implementation, sentences are tokenized using a pre-trained cased BERT tokenizer and fed into a BERT model transformer with a sequence classification head fine-tuned on the labeled training data. We also implement the BERTweet model, which shares its architecture with BERT but was trained on English tweets (Nguyen et al., 2020).

GPT-3: The GPT-3 architecture (Brown et al., 2020) introduced in 2020 achieved strong performance on many classic NLP tasks in a one-shot and few-shot setting. In other tasks, fine-tuning is required for good performance. We implement fine-tuning with both the Ada (least powerful) and Curie (second most powerful) engines. For GPT-3, fine-tuning involved using passages from the training data with the separator -> appended, with completion of either “sexist” or “not” depending on the label. For example, one row of the fine-tuning dataset looks like this:

prompt : “men without a women can think and be sucfull ->”,
completion : “sexist”

2.1.2 Performance

In accordance with the EDOS tasks guidelines, Macro F1-scores were used to evaluate each approach – see Table 1. Even without hyperparameter optimization and without the highest power engine, the fine-tuned GPT-3 model’s performance (0.854) is not particularly far from the top leaderboard performance (0.875).

<table>
<thead>
<tr>
<th>Model</th>
<th>Macro F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-LSTM</td>
<td>0.670</td>
</tr>
<tr>
<td>BERT</td>
<td>0.725</td>
</tr>
<tr>
<td>BERTweet</td>
<td>0.662</td>
</tr>
<tr>
<td>GPT-3-FT-Ada</td>
<td>0.830</td>
</tr>
<tr>
<td>GPT-3-FT-Curie</td>
<td>0.854</td>
</tr>
</tbody>
</table>

Table 1: Model performances on Task A binary classification.

2.2 Tasks B and C

2.2.1 Models

GPT-3: Given the high performance of a fine-tuned GPT-3 model using the Curie engine on Task A, we used the same architecture fine-tuned to the subcategories in B and C. Instead of completions “sexist” or “not,” we use the numbers corresponding to the subcategory. For task B, for instance, a row of the fine-tuning dataset looks like this:

prompt : “men without a women can think and be sucfull ->”,
completion : “3”

2.2.2 Performance

This approach yielded decent performance on the subcategory classification tasks, achieving Macro F1-Scores of 0.647 and 0.464 respectively (compared to leaderboard highs of 0.732 and 0.561).

3 Approaches to Explainability

In this section, we discuss and critique various theoretical approaches to explainability and interpretability, in particular focusing on the approach implicit in the EDOS’ use of sub-categories. This discussion motivates the counterfactual replacement analysis technique introduced in the section after.

It has been widely observed that there is very little agreement about what is meant by explainability and interpretability (Zhang et al., 2020; Lipton, 2016). In this paper, we use the two terms interchangeably to refer to the degree to which a models behavior can be understood. In other words, what logical decision rules (borrowing from Zhang et al., 2020) govern the relationship between a models inputs and outputs.

3.1 EDOS Approach

The EDOS task aims in part to provide an approach to explainability; the stated rationale for tasks B
and C is that classifying subcategories of sexism beyond binary detection “aids explainability in predictions.” The intuition here is simple and draws from the way we ordinarily justify our judgments about concepts like sexism:

JOHN DOW: Women are pigs.
JANE ROE: That is sexist.
JOHN DOW: Why is that sexist?
JANE ROE: It is dehumanizing, and derogates women by comparing them to non-human entities.

Roe’s explanation of why Dow’s statement is sexist corresponds neatly to a category (2. Derogation) and a vector (2.3 Dehumanising attacks and overt sexual objectification) in the EDOS taxonomy. The idea goes, then, that if we want to know why a model has predicted that a passage is sexist, we can understand by simply examining its category and fine-grained vector. We now move to discuss two characteristics of this sort of approach; that it is post-hoc, and that it is prescriptively taxonomic.

### 3.2 Post-Hoc Explainability

The EDOS approach to explainability is post-hoc; it asks a model to justify why a prediction has been made after the fact. Some post-hoc models are categorical (like EDOS’), while others might simply output a text explanation.

The issue with post-hoc approaches is that there is often a wide gulf between post-hoc justifications and the actual causal ‘reasoning’ used when determining whether or not the sentence in question is sexist. Consider the following ordinary interaction:

JAMES LO: I love my wife.
JENNY POE: Why do you love your wife?
JAMES LO: Because she is beautiful and kind.

Despite James’ ordinary assertion, the fact that James finds his wife to be beautiful and kind is not necessarily causally related or constitutive of him loving her. If James were to meet someone more beautiful and kind than his wife, it does not follow that he will or should love them instead. And James may continue to love his wife even if she were to become horribly disfigured.1 This example is illustrative of the fact that the actual mechanisms that our classification and decision-making function by are often quite different from what we reasonably report about them. This gap between reported post-hoc explanations and the actual mechanisms by which judgments are made has been widely observed in moral psychology (Haidt, 2001; Björklund et al., 2000).

While the above discussion refers to human explanations, it also applies to explainability in ML. To see this, it is helpful to consider what about post-hoc explanations are problematic. The problem with post-hoc explanations is not purely that they take place after judgments, but that they are not logically related to the actual decision procedure used in reaching the judgment. Thus, even if they are generated prior to the judgment, inaccurate yet superficially plausible explanations are common so long as they are logically disconnected from decision procedure. For example, the tendency to provide unfaithful explanations has been observed in chain-of-thought explanations generated by large language models prior to prediction (Turpin et al., 2023). Related problems inherent to post-hoc ML explainability by virtue of being logically disconnected from decision procedure have been observed in a variety of specific contexts (Bordt et al., 2022; Laugel et al., 2019).

Some have argued that the issues discussed above are inevitable in any method for explaining black-box models, and therefore we should pursue inherently interpretable models instead (Rudin, 2019); we are sympathetic to this concern, though we proceed with the assumption that genuine explainability for black-box models is possible.2

### 3.3 Prescriptive Taxonomic Explainability

The second characteristic of the EDOS approach is that it involves a prescriptive taxonomy derived from an extensive study of recent taxonomies of sexism.

However, it is worth noting that most complex and organically-developed concepts like sexism do not have the sort of strict and definite application conditions that this sort of taxonomy suggests. 3

---

1These issues are widely recognized in the philosophy of love. The former is referred to as the problem of trading up, and the latter the problem of inconstancy (Setiya, 2014).

2This section’s criticisms of the EDOS approach raise an obvious question: what if the fine-grained models classifications were used in the process of making binary classifications, rather than explaining them post-hoc? Put succinctly, to do so would be eschewing the problem of explainability for black-box models entirely in favor of inherent interpretability. In doing so we don’t gain any insight into the functioning of the black-box.

3In contrast, we can consider a stipulated concept like
This is widely observed in ML approaches to modeling human moral judgment, where morality is defined as much by rules as by the exceptions to them (Jin et al., 2022).

This issue is not related to the practical limitations of models. Even a model which perfectly classifies sexist remarks using the EDOS taxonomy is doing just that; classifying remarks using the EDOS taxonomy. This can be valuable and informative, but it only offers top-down explainability defined by a taxonomy that has already been set.

One of the most notable philosophical discussions of the difficulty of working with complex concepts is given by Ludwig Wittgenstein in his influential work *Philosophical Investigations*, where he argues that we can easily use the concept of “game” (e.g. board games, ball games, Olympic games) even though it lacks any core essence or application conditions (Wittgenstein and Anscombe, 1953).

Fortunately, Wittgenstein also provides a starting point for a more organic sort of interpretability in his understanding of meaning as use:

“For a large class of cases of the employment of the word ‘meaning’—though not for all—this word can be explained in this way: the meaning of a word is its use in the language” (Wittgenstein 43)

We proceed to develop our approach in the spirit of this observation. If meaning is defined by use, we can gain insight into how a model understands a concept by exploring in which of the cases (generated by replacement) the concept does and does not apply (i.e. which replacement sentences are labeled sexist or non-sexist).

4 Counterfactual Replacement Analysis

The motivation for our algorithm is to provide global explainability while avoiding the pitfalls of post-hoc and prescriptively taxonomic approaches. With this in mind, we advance a prototype for an approach where explainability is emergent from model behavior. We assume nothing ahead of time about the concept of sexism, and work to derive explainability only from the outputs of our black-box model.

In particular, our technique makes minimal perturbations to inputs at the margins and examines in what cases and in what ways these changes affect the application of the concept in question.

4.1 Similar Approaches

Applying minor perturbations to datasets is not new. Within the realm of sexism detection and classification, perturbations have been used to augment training datasets, boosting performance and improving generalizability on out-of-domain examples (Samory et al., 2021; Sen et al., 2021; Vidgen et al., 2021; Kirk et al., 2022). More generally, it is the core mechanism of the Local Interpretable Model-Agnostic Explanations (LIME) approach to interpretability, which addresses the problem of local explanations for model behavior on specific instances (Ribeiro et al., 2016).

4.2 Our Method

In this section, we walk through an early version of a novel technique that provides an angle for interpreting, understanding, and explaining text classification models.

This algorithm aims to analyze why a post is in a positive class (e.g. “sexist”) or a negative class (e.g. “non-sexist”), or more precisely, why a model predicts the post to be in one class or the other. The technique analyzes the impact (on the model’s predictions) of replacing a salient word with semantically similar words, in the process shedding light on the role the salient word plays in the classification of the examined post. 

Our analysis focuses on sentences that are both labeled sexist (according to the human labelers) and predicted sexist (according to the fine-tuned GPT-3 model). In other words, it is focused on true positives. Of the 2800 sentences in the testing data, 511 fit in this category.

4.2.1 Seed Word Groupings

We begin by grouping salient words in the corpus by semantic similarity. In theory, this process can and should be done algorithmically. In this paper, however, we focus on a singular group of semantically similar words chosen explicitly by the researchers: female-coded subject words, i.e.: women, woman, female, girls, females, ladies, and girl. This grouping is also desirable because of...
its frequency in the dataset provided. Of the 511 true positive sentences, 275 contained a word in the grouping.

4.2.2 Replacement Words

We now generate a set of “replacement words” – words semantically similar to the seed word grouping $G_h$. The aggregate value $v$ of a candidate replacement word $r$ is given by:

$$v(r) = \sum_{w \in G_h} \max(1 - \text{dist}(r, w), 0)$$

Candidate replacement words are only considered if they appear more than 400 times in the embedding corpus. The 50 replacement words with the highest value are chosen – see Table 2. Unsurprisingly, all words in the word grouping $G_h$ are also present in the set of replacement words.

<table>
<thead>
<tr>
<th>female</th>
<th>woman</th>
<th>women</th>
<th>girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>girl</td>
<td>wives</td>
<td>females</td>
<td>men</td>
</tr>
<tr>
<td>ladies</td>
<td>girlfriend</td>
<td>wife</td>
<td>girlfriend</td>
</tr>
<tr>
<td>people</td>
<td>sex</td>
<td>users</td>
<td>male</td>
</tr>
<tr>
<td>gender</td>
<td>chicks</td>
<td>queen</td>
<td>adult</td>
</tr>
<tr>
<td>sexual</td>
<td>young</td>
<td>lady</td>
<td>boys</td>
</tr>
<tr>
<td>man</td>
<td>user</td>
<td>daughter</td>
<td>children</td>
</tr>
<tr>
<td>kids</td>
<td>feminists</td>
<td>human</td>
<td>feminine</td>
</tr>
<tr>
<td>genders</td>
<td>teen</td>
<td>chick</td>
<td>feminist</td>
</tr>
<tr>
<td>humans</td>
<td>sexy</td>
<td>workers</td>
<td>wise</td>
</tr>
<tr>
<td>redpill</td>
<td>sons</td>
<td>boy</td>
<td>boyfriend</td>
</tr>
<tr>
<td>baby</td>
<td>guys</td>
<td>kid</td>
<td>gay</td>
</tr>
<tr>
<td>bitches</td>
<td>student</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Generated replacement words.

Eleven hardcoded replacement names were also added for consideration: olivia, emma, sophia, shanice, letitia, noah, jacob, michael, pelosi, hillary and obama. The first eight names were chosen from lists of popular (white female, black female, and white male) names. The last three are names associated with prominent political figures. While they are not in the top semantically similar words to the initial examined group, they are similar to those words in many respects and meaningfully different in others, which makes them likely to be informative replacement words.

4.2.3 Replacement Probing

Recall, there are 275 posts that contain a word in the seed grouping $G_h$. For each post, the first occurrence of a word in $G_h$ is replaced with each of the other 60 replacement words. For example, for an initial post “and some women fuck for money. yes, some women are whores,” the bolded word women is substituted for a replacement word like “female,” “woman,” or “boys.” The assumption behind applying the same replacement across a variety of posts is that replacement (of the same seed word with the same replacement word) functions as a similar sort of perturbation across all passages; in this way, we get a sense of the models behavior in various cases where parallel counterfactual modifications have been made.

It is worth noting that our approach assumes that the language model utilized can meaningfully interpret perturbations that do not retain grammatical or logical coherence – i.e. “woman can’t drive” $\rightarrow$ “feminism can’t drive.” While we suspect language models are able to reasonably parse this sort of sentence despite its incoherence, future approaches might attempt to maintain coherence in the process of perturbations.

For each post, and for each replacement word, there is a resulting replacement post, which is evaluated with the black-box model to get the predicted probability that it is a positive class (sexist). The predicted probabilities can be thought of as forming a table which is 61 columns (number of replacement words) by 275 rows (number of relevant sentences) – see Table 3.

The information expressed in this table constitutes the substance of the interpretation technique; it contains all the information we will use in the analysis. The issue is that there is too much information contained to be interpretable. Thus, the outputs of replacement probing are processed to condense information and enable meaningful interpretability.

4.2.4 Clustering Replacement Words

Each of the 61 replacement words has a distinct behavior given by the model outputs stored in its column in Table 3. Clustering replacement words based on their behavior clarifies the results of replacement probing.

A spectral clustering algorithm was used to group the different replacement words – see Table 4. Spectral clustering is ideal for picking up on datasets with convex or non-complex structures and allows for variation in density between groups. While spectral clustering does not guarantee hierarchical clusters, the clusters found were, in fact, hierarchical for $n$ from 2 to 7. The clusters are
Table 3: Table replacement scores

<table>
<thead>
<tr>
<th>Sentence</th>
<th><strong>female</strong></th>
<th><strong>woman</strong></th>
<th><strong>users</strong></th>
<th><strong>male</strong></th>
<th><strong>obama</strong></th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>it’s different if you’re a woman. spreading your legs just gives you sex</td>
<td>0.998</td>
<td>0.999</td>
<td>0.079</td>
<td>0.690</td>
<td>0.015</td>
<td>...</td>
</tr>
<tr>
<td>fuck those eyebrows. god i hate women.</td>
<td>0.999</td>
<td>0.999</td>
<td>0.000</td>
<td>0.105</td>
<td>0.000</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 4: Results of spectral clustering replacement words with 7 clusters labeled with a descriptor of the contents of the grouping.

<table>
<thead>
<tr>
<th>#</th>
<th>Description</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Sexuality</td>
<td>sex, sexual, sexy</td>
</tr>
<tr>
<td>1</td>
<td>Relationships</td>
<td>girlfriends, wife, girlfriend, daughter</td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
<td>gender, genders</td>
</tr>
<tr>
<td>3</td>
<td>Female Names</td>
<td>queen, olivia, emma, sophia, shanice, letitia</td>
</tr>
<tr>
<td>4</td>
<td>Women</td>
<td>girls, girl, wives, ladies, chicks, lady, chick</td>
</tr>
<tr>
<td>5</td>
<td>Men</td>
<td>men, people, users, male, adult, young, boys, man, user, children, kids, human, teen, humans, workers, wise, redpill, sons, boy, boyfriend, baby, guys, kid, gay, student, noah, jacob, michael, pelosi, hillary, obama</td>
</tr>
<tr>
<td>6</td>
<td>Fem/Feminism</td>
<td>female, woman, women, females, feminists, femin-nine, feminist, bitches</td>
</tr>
</tbody>
</table>

While the resulting clusters in table 4 intuitively resemble clusters based on semantic meaning, it is important to note that we are grouping by replacement behavior, not semantic meaning. To clarify this point, consider the inclusion of “feminists”, “feminist” and “bitches” in cluster #6 with 4 of the six words in the initial word grouping. Replacement cluster #6 is somewhat unusual: the mean predicted probability sexist for replacements from it was 0.969, while predicted probabilities for other clusters ranged from 0.562 to 0.897. This isn’t surprising. Because the algorithm only examines true positive cases (both predicted to be and actually sexist), we should expect the predicted probabilities for replacements using the replacement word cluster containing most of the initial word grouping to remain close to 1. The inclusion of words not in the initial word grouping (like “feminists”) in cluster #6 is due to them having a similar behavior as replacement words; in other words, replacing an initial word grouping word with “bitches” (for example) almost always results in a post that is (unsurprisingly) still predicted sexist. After we have grouped the 61 replacement words into 7 replacement clusters, we take the grouped mean probability for each post across the replacement words in its cluster, resulting in a table with 7 columns and 275 rows. In other words, for a post $p$ and replacement word cluster $c$, replacement score is given by:

$$score(p, c) = \frac{1}{|c|} \sum_{r \in c} M(p, r)$$

Where $M(p, r)$ denotes the model’s predicted probability that $p$ is sexist when its first word contained in the seed grouping is replaced by $r$.

4.2.5 Passage Clustering

Having reduced the dimensions of the replacement word axis from 61 to 7 via clustering, we then move to the other axis. We similarly cluster the 275 passages based on their replacement behavior with each of the 7 replacement word clusters. We use spectral clustering for this axis as well.\(^5\) See Table 5.

4.2.6 Interpretation

The proposed counterfactual replacement analysis takes a set of passages that contain a word in a given seed group of salient and semantically similar words, and groups them by their replacement behavior with respect to replacement words that

\(^5\)We think that there may be strong reasons to explore other clustering algorithms. They were not explored in this paper purely because of the time constraints of the competition.
are themselves grouped. The result is a table like 5.

We make the following observations based on the counterfactual replacement results.

1. Replacement word clusters have a (largely) tiered hierarchical structure. In other words, the sentences where (e.g.) replacement with a sexuality-related word is predicted sexist are simply the sentences where replacement with a female name is sexist, plus passage clusters 2 and 3. And sentences, where replacement with the gender-related word are predicted sexist, are the same as those for sexuality, plus cluster 4. In less technical terms, the model’s understanding of word categories with respect to sexism is one-dimensional. Reference to men is less sexist than a reference to female names, and a reference to female names is less sexist than a reference to sexuality, and so on.

2. About half of the sentences (passage cluster 0) are predicted to be sexist regardless of what the first female-coded subject word is replaced with. This might be because the reference to the female gender is communicated elsewhere (for instance in a pronoun like ‘her’), or because the reference to the female gender is not necessary for the sentence to be sexist.

3. We may want to take a closer look at the passages in cluster 3, which bucks the tiered trend observed earlier. Despite having some of the lower mean replacement scores in the categories “gender,” “women” and “women and feminism,” its mean replacement score for “sexuality” is comparatively high.

These are only a handful of observations, and more could be made by digging into the content of the passage clusters themselves.

5 Conclusion and Future Work

In this paper, we began by implementing a variety of approaches to the problem of classifying sexist language, and find that a fine-tuned GPT-3 model is able to achieve good results without any optimization or fine-tuning. We then turn to the problem of interpretation and argue that EDOS’ focus on subcategory classification can only give us so much interpretability. Instead, we introduce a new counterfactual replacement technique for understanding and evaluating text classification models working with complex concepts. Finally, we show how the results of this technique can be understood and interpreted.

It is worth noting that our introduction counterfactual replacement analysis (4) does not attempt to offer a robust and final approach to global explainability; instead, we offer it as a prototype and precursor to further work that aims at explainability emergent from behavior in counterfactual scenarios.

Future work on this technique should focus on 1) defining a rigorous procedure for identifying the initial seed word grouping, 2) comparing and optimizing different clustering algorithms for dimensionality reduction of both passages and replacement words, 3) investigating clustering algorithms that allow for the investigation of outliers, and 4) determining whether the technique’s output can be improved such that interpretation is even easier.
Acknowledgements

We would like to thank the two anonymous reviewers from SemEval 2023 who provided detailed and thoughtful feedback on the first version of this paper. Thank you also to James Dobson at Dartmouth College for his guidance in the early stages of research.

References

Jacobellis v. ohio.


Han Liu, Fatima Chiroma, and Ella Haig. 2018. Identification and classification of misogynous tweets using multi-classifier fusion.


Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.

A Appendix

Table 6 provides the details for architecture and parameters used.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bi-LSTM</strong></td>
<td>128-dimension embedding layer</td>
</tr>
<tr>
<td></td>
<td>Bidirectional LSTM with hidden size 64</td>
</tr>
<tr>
<td></td>
<td>Attention layer</td>
</tr>
<tr>
<td></td>
<td>Linear layer outputting size 2 out sample</td>
</tr>
<tr>
<td></td>
<td>10 epochs</td>
</tr>
<tr>
<td></td>
<td>1e-2 learning rate</td>
</tr>
<tr>
<td><strong>BERT/BERTweet</strong></td>
<td>128 batch size</td>
</tr>
<tr>
<td></td>
<td>5e-5 learning rate</td>
</tr>
<tr>
<td><strong>GPT-3</strong></td>
<td>0 temperature</td>
</tr>
</tbody>
</table>