Reproducing X-Class: Weakly-Supervised Hate Speech Classification

Iris Chen
George Mason University
ichen6@gmu.edu

Christophe Leung
George Mason University
cleung5@gmu.edu

Kelvin Lu
George Mason University
klu21@gmu.edu

1 Introduction

1.1 Task / Research Question Description

The purpose of this paper is to analyze the effectiveness of using a weakly-supervised classifier, X-Class, to classify hate speech on the Internet. Our goal is to reproduce the results found in (Jin et al., 2023). Moving forward, we aim to analyze the performance of a weakly-supervised X-Class model, as well as how the model would respond to multilinguality and erroneous training data. In this context, erroneous training data would be sentences that include noise, typos, grammar mistakes, and ambiguity.

(Jin et al., 2023) was able to find comparable accuracies with fully-supervised and fine-tuned training models. In their research, it was noted that the fully-supervised and fine-tuned training models, namely BERT and SVM, was able to outperform the weakly-supervised X-Class Classifier, however, not by an overwhelming amount. To reproduce this study, we will implement X-Class and BERT, train both models on the available dataset that (Jin et al., 2023) used, then compare our X-Class results with our BERT results. To decide whether we have been successful in reproducing the findings of (Jin et al., 2023), we will compare the difference of our BERT and X-Class performance metrics against the difference of (Jin et al., 2023)'s X-Class and BERT performance metrics.

1.2 Motivation and Limitations of Existing Work

Hate speech identification and classification is an active area of research in the field of natural language processing. There are many papers in ACL2023 that tackle the same problem from different perspectives, with some papers providing more labeled datasets and others that introduce new methods for higher classification accuracies. The reason for the focus on hate speech classification is because of the growing prevalence of it on the public Internet. The effect of hate speech can be seen in the rise of extremists groups and Internet-related harassment as evident in current events.

As traditional, strongly-supervised classification methods face issues with poor generalization over unseen datasets as well as datasets labeled using different taxonomies, our chosen paper (Jin et al., 2023) uniquely experiments with the effectiveness of extremely weak supervision on hate speech classification using the X-Class Classifier. In their research, it has been observed that the X-Class Classifier achieves a comparable performance to the fully-supervised baselines. As thus, the weakly-supervised text classifier presents a viable way to train and classify hate speech with only a few labeled seed words and without the use of any labeled documents. Since the need for a large labeled dataset is a limitation of prior work, being able to require less training data by eliminating the need for a labeled dataset is a beneficial advancement in the field of hate speech classification.

The limitations of (Jin et al., 2023)'s paper is the lack of multilinguality and robustness testing. As mentioned earlier, hate speech identification and classification is an active field of study. It is also important to note that a shortcoming of this study is that it focuses on classification rather than identification. Thus, if provided with non-hate speech text, it may label it in unexpected ways as this model is only trained to classify different categories of hate speech.

1.3 Proposed Approach

To reproduce the findings of (Jin et al., 2023), the preliminary approach is to use the XClass Clas-

Compiled 2023–11–04 03:56:04 UTC

sifier code as published in (Wang et al., 2021) and run our own hate speech classification. Once the repository is cloned and is running properly on our system, we will train our model using the cited dataset from the paper (Sap et al., 2020). To compare the XClass Classifier against a strongly-supervised classifier, we will train and classify the same datasets using BERT. Our implementation of BERT will be an adapted and largely re-written version the BERT pipeline used in the CS678 class assignments. Afterwards, we will examine the accuracy, precision, recall, and F1 scores of our X-Class model and compare them with the corresponding performance metrics of the BERT model. To determine whether our reproduction was successful, we will compare the difference between our X-Class and BERT performance metrics with the difference in performance metrics that (Jin et al., 2023) found between their implementation of BERT and X-Class.

If we find that the performance results of our implementation of BERT and X-Class are not comparable to the results in the paper, we will further examine the results as well as both ours and the cited implementations to understand the root cause of the variations. We will also conduct an error analysis to identify the discrepancies between the datasets and differences in any underlying processes. It should be noted that there is a significant amount of ambiguity in (Jin et al., 2023)—for example the data prepossessing techniques, updates to the dataset, as well as certain implementation details that were not mentioned to name a few—that may be a cause for some noticeable discrepancies in the results.

1.4 Potential Challenges and Mitigations

Aside from the time constraint of reproducing the results of this paper within 2-3 weeks, the main difficulty in this task is to be able to run the provided code and modify it in a way that allows us to reproduce the experiments conducted in the paper. It is important to note that the paper (Jin et al., 2023) itself does *not* provide any code nor datasets, but that it cites the use of the XClass Classifier from an external source, (Wang et al., 2021). Likewise, the final datasets used in (Jin et al., 2023)'s paper were also not provided and instead only linked to the data source. As thus, we may encounter unforeseen problems as we recreate the entire training and testing environment to house and run the XClass Classifier. Re-preprocessing the data without having a clear set of instructions from (Jin et al., 2023)'s paper would introduce yet another variable in the results. Another important point of note is that the provided XClass code is implemented in a very messily way, making it extremely challenging to audit its implementation.

In the event that the paper cannot be followed and the official implementation of X-Class cannot be used, we may consider only implementing the SVM and BERT models that the X-Class is compared to, training them on the available dataset that the paper uses, and try to recreate comparable performance results that the paper finds for SVM and BERT. The reason SVM or BERT models may be more doable is due to the fact that they are both used and implemented using scikit-learn and huggingface libraries. As these two libraries are widely used, they are much more well-maintained and documented, making remedies much easier.

2 Related Work

In this section we will examine studies that use concepts directly related to the paper we are trying to reproduce or concepts that would help us further the results of our model. By examining these related works, we hope to have a better understanding of how to implement or improve the original paper.

The paper we are doing a reproducibility study on uses the official implementation of X-Class provided by (Wang et al., 2021). This paper contributes the original and official development and implementation of X-Class. X-Class is a way to explore "extremely weak supervision" through representation learning. This novel method first generates class-oriented document representations, then forms document-class clusters before training the model. Essentially, through clustering, pseudo training data can be generated for use with a traditional neural language model without the need for manually-labeled (supervised) training examples. This paper is an important source as we may need to refer to it when encountering issues with our implementation of the X-Class code it provides.

(Antypas and Camacho-Collados, 2023) provides numerous datasets of hate speech and performs a large-scale cross-dataset comparison. It concludes that combining these datasets results

in a greater robustness in the classification models used. An important takeaway from (Antypas and Camacho-Collados, 2023) is that it shows how combining datasets contributes to the robustness of hate speech detection models. With the findings of this paper in mind, we may consider incorporating the datasets in this paper for cross-dataset classification to further improve our X-Class model for robustness.

(Goldzycher et al., 2023) evaluates whether using a limited amount of labeled data (in a target language) can perform well with Natural Language Inference (NLI) models. This paper focuses on the difficulty of classifying with non-English datasets which do not have as much data to be trained on as English datasets do. This paper finds that NLI fine-tuning leads to strong improvements to hate speech detection in non-English datasets, though not to the same extent as English. This concept of having limited labeled training data in a certain language is identical to the weakly-supervised classifier. To implement multilinguality with the model we create for this reproducibility study, we may consider using NLI fine-tuning to improve non-English classification.

Using a different model, (Yoder et al., 2023) accomplishes a similar goal of hate speech detection with a weakly supervised classifier. The main differences between this paper and ours is that (Yoder et al., 2023) focuses mostly on hate speech relating to white supremacy and that this paper uses DistilBERT rather than X-Class to do its classification. (Yoder et al., 2023) finds that introducing neutral and anti-racist data into their training data helped in mitigating bias in their model. If we find during error analysis that our model has bias, then it may be helpful to introduce more non-hate speech into our training data.

3 Experiments

3.1 Datasets

The dataset we will be using is the set present in (Jin et al., 2023) as linked below. Originally, this paper cited the use of two datasets, the Waseem dataset (Waseem, 2016) and the SBIC dataset (Sap et al., 2020), both of which are popular hate speech sources with regards to their categories. However, only the SBIC dataset was available for use in our experiments. The Waseem dataset repository does not include the actual text data, but rather a list of post IDs to the social media platform *X* (formerly known as *Twitter*). As these post IDs are only accessible through X's API, which at the time of writing costs $100 USD per month, retrieving the *Waseem* dataset for use in our reproducibility study is beyond the budget available.

`https://maartensap.com/social-bias-frames/`

The SBIC dataset, as mentioned in (Jin et al., 2023), does not have a predefined hate speech taxonomy. To address this problem, we use the "targeted minority" of each post as the label. This paper chooses the following six most common target minorities as labels for the SBIC dataset: Women, LGBT, Black, Jewish, Muslim, and Asian. However, even though these labels are specified, it is not defined exactly what targeted minority labels fall into these categories. For example, there is no LGBT labels, but there are labels such as "trans women" and "lesbian women." Similarly, it is not specified whether labels such as "Chinese" also fall within "Asian." This leads to some discrepancies between our curated datasets, as seen in Figure 1. Additional discrepancies could also be due to the fact that it is not explicitly stated what is done when there are multiple groups in the targeted minorities.

For our curated version of the SBIC dataset, we have classified each of the 6 subsets as follows: If any instances of "gay", "lesbian", "trans", "bisexual", or "asexual" appear in the label, then it is labeled as LGBT. Then, for the remaining labels, a post is given the label if the text "women", "black", "jewish", "muslim", or "asian" appears in the targeted minorities in the given priority. The reason for this is because we believe it would be more likely that minorities including those LGBT labels such as "lesbian women" are more targeted at the LGBT community than it is meant to be targeted at women. The remaining priority order was chosen simply as (Jin et al., 2023) displayed it.

The train/dev/test split will be the specified split in (Jin et al., 2023) for the SBIC dataset where the training is 75% of the data, and the remaining 25% is split evenly between dev and test (12.5% for each). There is already a predefined split of the data as (Sap et al., 2020) provides the dataset as six files, two for test, two for train, and two for the dev splits. Which of the two test/train/dev datasets (Jin et al., 2023) uses is not specified, but it is evident that the paper uses the smaller size datasets (SBIC.v2.agg.train.csv, SBIC.v2.agg.dev.csv, and

SBIC.v2.agg.test.csv), based on their table 2 in section 4.2 where their total was around 7 thousand samples. Meanwhile, the other available datasets (SBIC.v2.train.csv, SBIC.v2.dev.csv, and SBIC.v2.test.csv) have tens of thousands of samples. It should be noted that (Jin et al., 2023) explicitly states that they use the default data split that (Sap et al., 2020) provides. However, in our data validation checks, we found that the default splits provided in the SBIC dataset did not align with the cited 75%/12.5%/12.5% split proportions. Thus, for our experiment, we have chosen the "agg" version the data, combined the pre-split data ourselves and re-split the data to a 75%/12.5%/12.5% split using an identical shuffle for both the XClass and BERT classifiers.

Table 1: Dataset comparison to Jin et al. 2023

	Category	#Train	#Test
Orig.	Women	2,594	351
	Black folks	2,512	576
	Jewish folks	847	207
	LGBT folks	490	53
	Muslim folks	412	85
	Asian folks	224	34
Ours	Women	2,800	401
	Black folks	2,803	644
	Jewish folks	930	205
	LGBT folks	640	96
	Muslim folks	469	87
	Asian folks	261	44

Figure 1: A table comparing the number of training and test between our version of data curated from SBIC and Jin et al. 2023's curated data. We are labeled as "Ours" and Jin et al. 2023 is labeled as "Orig."

(Jin et al., 2023) states that they use preprocessing methods following (Barbieri et al., 2020) where they (1) anonymize users by replacing references to users with "@user", (2) anonymize website links by replacing links with "http" (3) and removing emojis, which is not included in (Barbieri et al., 2020). The preprocessing code is provided in (Barbieri et al., 2020) and is linked below. However, The provided code for preprocessing did not accomplish what it claimed, with many edge cases not included that were prevalent in the SBIC dataset. Therefore, we were not able to utilize this code and instead developed our own preprocessing code. An additional preprocessing step we have included was (4) removing unknown tags in the SBIC data in the format "&#[six digits];" (eg. 📹). It is not specified nor intuitively recognizable what this tag is meant to represent, but we have decided to remove it for the sake of better understanding of our data. For tokenization we use the built-in tokenizer in the official X-Class implementation.

https://github.com/cardiffnlp/tweeteval#tweeteval-the-benchmark

3.2 Implementation

For our implementation, we utilized and modified the official weakly-supervised X-Class provided by (Wang et al., 2021) as this the resource that (Jin et al., 2023) had cited as using. For implementation of BERT, as mentioned earlier in section 1.3, we repurposed the code provided in HW2 of this class.

Linked below is our repository labeled "Our repository" and the official implementation of X-Class labeled "Official implementation of X-Class".

Our repository: https://github.com/ehpotsirhc/hatespeechdetect

Official implementation of X-Class: https://github.com/ZihanWangKi/XClass

3.3 Results

Results from our reproduction of X-Class and BERT with the preprocessed SBIC dataset are labeled as "Ours" in Figure 2. These results are compared with performance metrics from Section 4.4 of (Jin et al., 2023), recreated below with only the results of their BERT and X-Class on the SBIC dataset. Their results are labeled "Orig."

Table 2: SBIC Dataset

	Accuracy	Precision	Recall	F1
Orig				
BERT	95.7	94.2	95.1	94.6
XClass	79.8	74.0	81.8	74.8
Ours				
BERT	93.06	93.1	93.1	93.1
XClass	75.4	75.4	75.4	75.4

Figure 2: Accuracy, Precision, Recall, and F1 scores of our reproduced results on X-Class and BERT compared to (Jin et al., 2023)'s results for the SBIC Dataset.

3.4 Discussion

Aside from sensitivity to randomness, there are many reasons why there may be significant discrepancies between our results and that of (Jin et al., 2023). Since (Jin et al., 2023) did not directly provide their implementation and modifications of X-Class, preprocessed data, and overall pipeline, we needed to create it ourselves. There are many places where our implementation of the pipeline could differ from the original paper's pipeline. This section covers where some of these discrepancies may lie.

One of the issues we faced is implementing and debugging the provided X-Class repository. After first cloning the repository provided by (Wang et al., 2021) and running it with the provided datasets in the repo, we ran into errors.

As mentioned earlier in section 3.1, we had originally wanted to include both datasets that (Jin et al., 2023) used in their research. However, the Waseem dataset (Waseem, 2016) provided only X post IDs as opposed to text. Earlier this year, X (formerly Twitter) had began requiring $100 USD per month to access its API, which made the Waseem dataset no longer available for us to use.

Unlike with the Waseem dataset, issues we faced regarding the SBIC dataset (Sap et al., 2020) were due to ambiguity in (Jin et al., 2023). As mentioned in section 3.1, the authors did not specify their curation process of the data. Because of this, we were not able to recreate the exact dataset that (Jin et al., 2023) used in their study. Figure 1 shows the differences in our data count. Similarly, the data split of 75%/12%/12% as described by the authors is not exact based on the sum of their data in Figure 1. This ambiguity would be insignificant if not for the fact that we are trying to reproduce the study as closely as possible.

(Jin et al., 2023) states that they follow the preprocessing pipeline of (Barbieri et al., 2020) where they "including user mention anonymization and website links and emoji removal." However, (Barbieri et al., 2020) does not do emoji removal, and (Barbieri et al., 2020) does not specify whether they use the same tokenization that (Barbieri et al., 2020) uses. Additionally, it is not specified whether (Jin et al., 2023) removes the tags we had mentioned in section 3.1 that are uninterpretable and presumably not supposed to be in the post text. This is another example of ambiguity present in (Jin et al., 2023) that causes issues with making our reproducibility study as similar as possible.

Another discrepancy lies in section 4.3 of (Jin et al., 2023) where the authors state that they used the bert-based-uncased checkpoint to fine-tune their final classifier of X-Class. We are unclear as to what this means, as the use of a fully-supervised fine-tuned model in conjunction with X-Class seems to be in antithesis of the purpose of X-Class, being a weakly-supervised model.

3.5 Resources

The time cost of reproducing this study was the greatest challenge. As we were only provided a few weeks to reproduce this study, it does not compare to the resources and time that (Jin et al., 2023) may have been given. Aside from the time resource, we also required memory, computation, and development efforts. In this section we report the measurable resources used.

Table 3: Resources

Resource	Usage
Implementation Time	60 hours
Training & Tuning	10 hours
GPU	1 $\times$ 16 GBs

Figure 3: Measurable resource usage for time, memory, and computation.

3.6 Error Analysis

In this section we analyze instances where our model failed to classify samples with the correct label. Since we use two different datasets, we do this analysis with both of them.

Content Warning: Please note that as this study focuses on hate speech, the samples in the table may contain inappropriate, sensitive, and/or harmful content. The authors of the original paper and the students performing this reproducibility study do not support the use of this harmful language.

In figure 4, we see that our X-Class model wrongly classified the first sample's target minority as the Jewish community rather than the black community. Due to the nature of the way this sample was phrased, it introduces ambiguity that our model may not have been able to understand as there seems to be no direct terms that would indicate this sample as text targetting the black community. It should be noted that that the SBIC

dataset have disproportionate distributions of labels in their data. We recreate table 2 of section 4.2 in (Jin et al., 2023) as figure 1. With this distribution in mind, we may keep in mind that samples with significantly less training data (eg. Jewish folks) may be more likely to be falsely classified with a label with significantly more training data (eg. Black folks), which may have been the cause of the first mis-classification seen in figure 4.

Table 4: Failed Classification Examples: SBIC Dataset

Sample	True	Predicted
What do you call a sniper in the hood Pest? exterminator.	black	jewish
how do you break up a ten man gang rape. toss them a basketball.	black	women
If you still use Twitter or Facebook, you're a kike shill	jewish	asian

Figure 4: Incorrectly labeled samples from our model in the SBIC Dataset, their correct labels, and our incorrectly assigned labels.

In the second sample in Figure 4, we see that the correct targeted community is the black community, while the correct label is women. Looking at the text, we can understand that it would not be unreasonable to put this text under both minorities, as it shows disrespect to both. "Rape" is often a term used heavily in hate speech against women, so to see our classifier label the targeted minority to be women is understandable.

In the third sample in Figure 4, we see that our classifier mistakes a post targeting the Jewish community as targeting the Asian community instead. The slur used in this text should have been a significant indicator that this is targeting the Jewish community. Meanwhile, there seems to be no direct indicator that is discriminatory towards the Asian community. This is a curious misclassification, as the other two misclassifications seemed to have reasonable logic behind mistaking one label for another.

The original study by (Jin et al., 2023) has section 4.5 where it analyzes the difficulty of cross-dataset classification. It notes especially that going from the Waseem dataset, which only has 2 labels, to the SBIC dataset, which we've given 6 labels, has significantly lower performance metrics. This is a significant error analysis, however it would have been interesting to see similar examples as we provided above, where we can see exactly where the model went wrong. Our study does not incorporate cross-dataset classification, as the Waseem dataset was not available to us, and we chose to focus on properly implementing X-Class.

3.7 Implementation Auditing

The provided code for XClass is arguably written in a manner that made it very challenging to audit. There was minimal documentation on how the code should be run outside of the predefined parameters. To add, although the runtime parameters seem to allow for an evaluation-only mode, the parameters were riddled with bugs and it was impossible to run a test without running the training together, increasing the complexity and time necessary to run our testing data. It is also worth noting that the default implementation does not have any validation dataset as well. As XClass generates pseudo training data as the training data to the neural language model, the pseudo training data is tested against the original training data as the validation. Such a practice could affect the overall accuracy of the results.

4 Conclusion

In this paper we have made a reasonable attempt at a reproducibility study of (Jin et al., 2023), where they used the weakly-supervised X-Class classification mode developed by (Wang et al., 2021) to classify hate speech. To the best of our ability, we used the official implementation of X-Class, training it on the SBIC dataset (Sap et al., 2020), created our own preprocessing pipeline, and finally trained our X-Class model as well as BERT (adapted from Homework 2 of this class) to classify hate speech.

The performance metrics of our implementation of X-Class and BERT are comparable to the performance metric results from the X-Class and BERT in (Jin et al., 2023), and therefore would lead us to believe that this paper is reproducible. However it should be noted that we were not able to recreate one of the other contributions in this paper—the cross-dataset classification with X-Class—as we did not have access to the additional dataset (the Waseem Dataset (Waseem, 2016)). Despite this, we were able to recreate the comparable results between X-Class and BERT

with hate speech classification as the paper described as well as analyzing our results.

Regardless of whether we were able to reproduce all the contributions in (Jin et al., 2023), we are confident in further modifying our X-Class model to incorporate robustness and multilinguality in future studies.

References

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Janis Goldzycher, Moritz Preisig, Chantal Amrhein, and Gerold Schneider. 2023. Evaluating the effectiveness of natural language inference for hate speech detection in languages with limited labeled data. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 187–201, Toronto, Canada. Association for Computational Linguistics.

Yiping Jin, Leo Wanner, Vishakha Kadam, and Alexander Shvets. 2023. Towards weakly-supervised hate speech classification across datasets. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 42–59, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-class: Text classification with extremely weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3043–3053, Online. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Michael Yoder, Ahmad Diab, David Brown, and Kathleen Carley. 2023. A weakly supervised classifier and dataset of white supremacist language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 172–185, Toronto, Canada. Association for Computational Linguistics.