Multilinguality and Robustness with X-Class on Classifying Hate Speech

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 multilinguality and robustness of the weakly-supervised classifier X-Class to classify hate speech on the Internet. Previously, we had re-implemented and evaluated the X-Class classifier against BERT, on the SBIC dataset as (Jin et al., 2023) had done, and we were able to achieve comparable results. In this paper, we aim to analyze the performance of this weakly-supervised X-Class model on how it would respond to multilinguality and robustness testing. In this context, robustness testing is defined by (K et al., 2022) with Minimum Functionality Tests (MFT), Invariance Tests (INV), and Directional Expectation Tests (DIR).

(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, the research lacked any testing on model robustness and support for handling languages outside of English. In theory, the X-Class classifier should be extremely robust and can handle different languages, yet in practice our experiments show a polarizing result. While we were able to reproduce a similar set of results when comparing the base X-Class to our implementation of BERT, BERT's performance far outweighs the performance of X-Class when it comes to robustness and multilinguality.

For our robustness and multilinguality experiments, we will rewrite the X-Class classifier and BERT; using the (Sap et al., 2020) dataset, create translated versions of this dataset from English to Hindi, Japanese, Russian and Spanish; create very simple samples for MFT robustness training; create erroneous versions of this dataset by introducing typos, expanding and contracting contractions for INV robustness training using (K et al., 2022)'s `checklist` library; swapping target minorities and corresponding slurs for DIR robustness training; then finally analyze our X-Class results alongside the BERT results. We will study how the X-Class results are affected by multilinguality and erroneous training data by comparing it to our previous X-Class results as the baseline comparison before introducing the two new aspects. Similarly, to see how X-Class performs against BERT, we will compare that disparity with BERT's comparison before introducing new languages and erroneous data.

1.2 Why is this problem important?

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.

Tackling robustness and multilinguality in this field is important because English-speaking Internet users are not the only group of individuals that are targeted and experience hate speech. In fact, one may argue that those who do not speak English are more likely to be susceptible to being targeted minorities of hate speech. Robustness is an aspect that is already inherent in our data, as these samples were taken from online forums. The nature of taking these posts from real online users means that there is likely already grammar, spelling, and self-censoring present in the samples. However, we aim to introduce more robustness training by modifying the data further. This is important as users learn to self-censor more and more as seen with the current trend of replacing ”kill” with ”unalive.”

1.3 Motivation and Limitations of Existing Work

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.

The motivation of our proposed study is addressing these two of the issues mentioned above. By introducing multilinguality and robustness training, we can observe how the X-Class classifier performs and speculate on how to move forward and improve these two aspects.

1.4 Proposed Approach

Previously, we had produced comparative results to the findings of (Jin et al., 2023), where we had used the X-Class classifier code as published in (Wang et al., 2021) and run our own hate speech classification. We had trained our model using the cited dataset from the paper (Sap et al., 2020) and compared the X-Class classifier against a strongly-supervised classifier, BERT, once again trained on the same dataset. In our study to test X-Class's robustness and support for other languages, we intend to maintain the same training and testing structure used in the experiments from our previous study, but that the datasets will be modified. To elaborate, we will perturb our SBIC dataset using (K et al., 2022)'s `checklist` PyPI library to test the robustness of X-Class, as well as translate the same SBIC dataset into different languages for testing our X-Class classifier's resilience to other languages. The experiments will be trained and tested on both X-Class and BERT so that X-Class's performance can be baselined against those of BERT's.

1.5 Summary of Results

The introduction of perturbed data for our robustness experiments has caused the performance of the X-Class classifier and to BERT-based classifier deteriorate drastically, with the X-Class classifier performing worse in every test except MFT. On the surface, this suggests that the X-Class classifier is ineffective at handling erroneous data. To add insult to injury, the X-Class classifier would not accept any of our four translated datasets. A detailed study of the X-Class classifier did reveal why the X-Class classifier failed to classify our translated datasets at all and will be discussed in our *Experiments* section. The BERT-based classifier, while able to classify our translated datasets, also suffered from an approximate 10% reduction in overall performance. Suffice to summarize, the X-Class classifier, at its current state, is limited with handling any data that is outside of its tested, unmodified English dataset, while BERT managed to handle all of the different tests with varying results in performance.

2 Approach

2.1 Background

The concept of weakly-supervised classification may not be a standard concept for students of this class. (Wang et al., 2021) describes weakly-supervised classification as being able to classify text using training data that has only been included in a few labeled documents per class. In simpler terms, this means that a weakly-supervised classification model is trained on less labeled data than a fully-supervised model. Weakly-supervised classification models are useful in reducing the amount of human effort necessary to manually label data. Subsequently, such reductions in labor also enables us to train new models with greater efficiency, resulting the ability to train on larger datasets, as well as eliminate the possibility of introducing any human biases during the labeling process. In the X-Class classifier, the model is able to infer its own labels given only the training corpus and a predefined set of class labels. The reason for the development and analysis of X-Class is to create an extremely-weakly-supervised model that may be comparable to fully-supervised models, such as BERT.

2.2 The X-Class Classifier

The X-Class classifier can best be understood as a text classifier that requires no manual labeling at all. The classifier achieves this ability by taking a corpus of sentences, or in our case the social media posts, tokenizes the sentences to extract the vocabulary from each document, clusters each of these extracted words with one of the class labels that have been pre-defined, then uses these clustered words to pre-classify each sentence into one of the pre-defined classes. The process effectively curates a set of self-labeled data which is then used to train a language model. In the original X-Class classifier, the default language model used in the the final training is Hugging Face's `bert-base-uncased`. Aside from choosing the appropriate class labels for the corpus of text used, no human interaction is involved.

For clarity and modularity, our re-written implementation of X-Class separates the X-Class model into five stages:

(Stage 1) Tokenize, extract and distill a set of vocabulary from the raw text/corpus/posts.

(Stage 2) Associate the vocabulary with each of the pre-defined classes; build the class-oriented document representations by vectorizing each document based on the vocabulary that is associated with each class.

(Stage 3) Align/pre-classify each document to its nearest class.

(Stage 4) Filter for only the documents that are self-labeled with a high degree of confidence.

(Stage 5) Train a language model using the self-labeled data (ie. BERT).

2.3 Approach, Motivations and Intuitions

In order to study X-Class's model robustness and resilience to the handling of other languages, translation and perturbation modifications will be performed on the *cleaned and preprocessed* training and testing dataset that was used in our baseline experiments ((Sap et al., 2020)).

For robustness testing, we will modify the dataset using (K et al., 2022)'s `checklist` library to introduce erroneous data to the dataset. The erroneous data for INV robustness testing has a version of the dataset with 97% typos, a version of the dataset where all contractions have been expanded (26% of the data changed), and a version of the dataset where all potential contractions are contracted (12% of the data changed). For the simple samples for the MFT testing, we created simple statements in the format of "I hate [target_minority]", and "I dislike [target_minority1], but I hate [target_minority2] more" to see if the model accurately classifies the target minorities. Finally for DIR robustness testing, 40% of the data was modified, swapping the target minorities and swapping the slurs related to those targets. This was done by scraping online resources to create an exhaustive list of slurs for each target minority, then trimming down this list of slurs by removing any slurs that did not appear already appear in the dataset, and finally swapping the target minorities and slurs when they appear in the dataset.

For multilinguality, we will introduce different languages into our training and testing dataset. As our original dataset, (Sap et al., 2020), is an English-only dataset, we will translate the entire dataset to Hindi, Japanese, Russian and Spanish using the `googletrans` PyPI package.

Using the modified datasets, models can then be trained using the X-Class classifier, as well as the BERT-based classifier as our evaluation baseline. While our implementation of BERT continues to be an adapted and largely re-written version of the BERT pipeline used in the CS678 class assignments, the original X-Class classifier that was provided has to be completely rewritten due to its unmanageable disorganization. The provided code for X-Class is arguably written in a manner that made it very challenging to understand and use. The included documentation on how the code should be run outside of the predefined parameters is minimal. 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 the testing data.

In order to perform a uniform comparison between the two models, we need to be able to understand thoroughly how the different components of X-Class interacts with each other, as well as be able to compartmentally modify select components to aid in our testing variations. As thus, the X-Class classifier was rewritten in a modular way to help us better understand what the underlying X-Class classifier code is doing, as well as streamline certain processes. Because the key idea of the X-Class classifier is to label its own dataset through methods of inference and clustering, then use this self-labeled dataset to train a *language model*, our re-written X-Class classifier can be separated into five distinct stages for its training process, of which the final stage reuses our implementation of the BERT-based classifier pipeline as *that* language model the X-Class classifier trains in its last step.

For evaluations, the metrics used in our previous paper—accuracy, precision, recall, and F1 scores—will once again be used as the basis of our robustness and multilinguality comparison to our BERT-based tests. The results from the robustness and multilinguality experiments will also be compared against the baseline tests (prior to introducing robustness and multilinguality) from our previous study.

Linked below is our code repository housing the code we used for our experiments. In it contains our modularized implementation of the original X-Class classifier, as well as our adapted implementation of the BERT-based classifier. A link to the original X-Class classifier has also been provided for reference purposes.

This Paper's Code Repository
https://github.com/ehpotsirhc/hatespeechdetect

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

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	4,024	401
	Black folks	4,080	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. &#128249;). 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

Moving forward into introducing multilinguality, we utilized the Google Translate API (PyPI `googletrans`) in order to translate our entire dataset to four languages, Hindi, Japanese, Russian and Spanish. The reason why we chose these target languages is because they were among the most popular languages and also had a good variation between the language syntax (ie. Japanese has no spaces; Spanish separates words into masculine forms and feminine forms).

To introduce erroneous data to evaluate robustness with regards to MFT, INV, and DIR, we modified our entire dataset in the following ways: for MFT, creating simple sample posts in the format of "I hate [target_minority]", and "I dislike [target_minority1], but I hate [target_minority2] more"; for INV, modifying the existing dataset by introducing typos and expanding and contracting contractions; and for DIR, modifying the existing dataset by swapping the target minorities and slurs within the samples.

3.2 Baseline Methods

The baseline we are comparing our methods to are the results from our previous paper. These results can be seen in Figure 1 in Section 1.5. We were able to achieve these results by implementing the official X-Class repository, preprocessing the SBIC dataset as described in the previous section, and training the classifier. Similarly, we used our adapted BERT implementation, trained it on the preprocessed SBIC dataset, and compared those results to X-Class. BERT will still be utilized here to compare how X-Class reacts to multilinguality and erroneous data and how BERT reacts to multilinguality and erroneous data.

3.3 Performance Metrics

The metrics selected for the basis of our performance evaluation will remain the same as our previous paper, which aligns with those used in the original paper (Jin et al., 2023). These metrics being accuracy, precision, recall, and F1. The accuracy is calculated by finding the percentage of total correctly classified labels over all the samples. Precision is calculated by finding the number of correct positive classifications over all positively classified labels, meaning that in this metric true negatives are not included. Recall is calculated by finding the number of correctly classified positive labels over the real positive labels. Finally, F1 is the harmonic mean between precision and recall, calculated by doubling the product of precision and recall, then dividing by the sum of precision and recall.

To calculate these metrics, we utilized the following `scikit-learn` libraries: `sklearn.metrics.accuracy_score`, `sklearn.metrics.precision_score`, `sklearn.metrics.recall_score`, `sklearn.metrics.f1_score`

3.4 Results and Figures

Model performance results from our experiments could be found in Figure 2. Our tests without any data perturbations nor change in language forms our baseline results, labeled “Baseline”. Our robustness tests are primarily comprised of the Minimum Functionality Tests (MFT), Invariance Test (INV), and Directional Expectation Tests (DIR). Our multilingual tests are labeled ”Multi-lang”. As there are multiple tests run for INV and MFT, as well as multiple languages for the multilingual tests, these test results are averaged and reported as one aggregate number in the results table below. As a precaution, we have ensured that the test results we are taking an average of are close enough to each other that taking the average of the tests would provide a representative result as opposed to an inaccurately skewed result.

Table 2: X-Class vs BERT

	Acc	**Prec**	**Rec**	**F1**
Baseline				
BERT	92.02	88.59	86.71	87.40
X-Class	84.18	78.52	84.02	79.28
MFT				
BERT	42.96	45.36	42.10	40.43
X-Class	45.65	47.40	41.32	40.69
INV				
BERT	91.35	87.61	83.84	85.34
X-Class	82.94	78.74	81.96	77.52
DIR				
BERT	91.28	89.70	91.51	90.38
X-Class	71.49	72.08	74.65	66.87
Multi-lang				
BERT	79.66	72.22	72.97	71.28
X-Class	n/a	n/a	n/a	n/a

Figure 2: Accuracy, Precision, Recall, and F1 scores of our reproduced results on X-Class and BERT before and after introducing multilinguality and robustness training.

3.5 Discussion on Results

From our experiments, the X-Class classifier has a comparable performance to the base BERT in our baseline tests as established in our previous paper. However, it can be observed that the performance of the X-Class classifier suffers slightly when erroneous data is introduced, deteriorates significantly when processing MFT-perturbed data, and finally incapable of classifying any of our translated datasets in the languages Hindi, Japanese, Russian and Spanish. On the contrary, the BERT-based classifier maintained a competitive performance throughout all experiments, exhibiting only a noticeable decrease in performance when performing the MFT robustness test and processing the translated languages.

In theory, the X-Class classifier should be fairly robust in handling perturbed sentences. However, our experiments showed that when it came to the Minimum Functionality Tests (MFT) and Directional Expectation Tests (DIR), the X-Class classifier performed significantly worse than the baseline tests. For MFT, X-Class performed poorly because X-Class’s vocabulary-clustering process takes into account only very limited amounts of semantic information. Because much of our MFTs are riddled with racial slurs, a type of word that requires contextual knowledge of which X-Class lacks, the X-Class classifier is unable to properly classify an MFT. Another reason is likely because the language of the simple samples we created altered the core definition of the original sentences (simplified it) so drastically that the X-Class model simply had trouble classifying the the sentence, causing the accuracy to dip below 50%. Very rarely will there be a post that blatantly states the poster individual's hate for a target minority, so the model likely had difficulty classifying those examples. For the INV and DIR tests, X-Class performed reasonably but not as well as BERT. This is likely due to the language and complexity of swapping out the slurs and target minorities. Not only are some slurs not as common as others, but the fact that some slurs do not always carry between sentences possibly confused the X-Class model.

Similarly, X-Class should in theory perform well when classifying text in a language other than English. In practice however, the classifier is not able to classify any of the translated text because it failed at Stage 2 of the X-Class classifier. Stage 2 of X-Class is when the vocabulary is associated with each of the pre-defined classes. This means that the class labels need to be derivable from the sentences, or in other words, the pre-defined class labels must be of words that exist in at least one of the documents in order for the vocabulary-to-class association to work. This requirement created several limitations that then became a barrier for the X-Class classifier to handle translated text:

- Class names must be in the same language as the main text
- One must have prior knowledge of the corpus before class labels can be defined
- Some languages use multiple variations of the same word, making it challenging to classify sentences into discrete class bins

A key observation is that our translation from English to Spanish did not translate the term "LGBT" to Spanish, hence the XClass model was unable to cluster sentences to the "LGBT" class label because none of the Spanish sentences contained the term "LGBT". Even when the class labels are translated into a different language, there is not always a 1-to-1 mapping between a word in English and a comparable word(s) in another language. For example, our translation package translated the English term "Asian" to "Asiático"(masculine) and "Asiática" (feminine). The English class name "Asian" could then only be translated into two class names "Asiático" and "Asiática", braking our class structure. Manually modifying the label to "Asiátic", the common characters of the two word conjugations, would also not work as "Asiátic" is not a valid vocabulary word in the training corpus.

Another multilinguality observation is that this the X-Class model would struggle with handling East Asian languages such as Chinese, Japanese and Korean. This is because during the vocabulary-class association process, the X-Class classifier needs to tokenize the sentences in order to compute the vocabulary-class associations. Because this tokenization process involves no language models, it is almost impossible to tokenize the East Asian languages due to the lack of spaces within the text (each character or combination of characters carry a meaning as opposed to just a pronunciation in English). This failure to tokenize the East Asian languages would preclude the X-Class classifier from operating on its subsequent stages for such languages.

It is crucial to note that due to our limited language abilities, we could only use a translator to translate the corpus and the class labels. As outlined above, these translations created a dataset that could not be used with the X-Class classifier. However, due to the way the X-Class classifier is setup, it is likely possible that X-Class would work on a corpus that is natively available in another language coupled by class labels curated by a native speaker of the foreign-language corpus.

3.6 Analysis of Classification Errors

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.

Listed in Figure 4 is an example of some perturbations that were used in our robustness testing. The first example is the original, and the subsequent examples are the DIR, INV and MFT examples. The word "pancake" is a racial slur that refers to Asians. For MFT, as explained in the previous section (see section *Discussion on Results*), X-Class has very limited semantic knowledge and hence could not understand any racial slurs. Such a limitation would also carry over to some of our DIR tests. In our DIR example listed, X-Class clearly prioritized the keyword ”she” for the classification as it had no knowledge that ”pancake” was a racial slur referring to Asians. Perhaps a larger dataset full of racial slurs could remedy this shortcoming but in our dataset, the examples contained within was not sufficient to train the X-Class classifier of any racial slurs. On the contrary, the X-Class classifier is relatively resilient to typos as evident in our example, as a typo here or there would not affect the overall vocabulary clustering process of the entire corpus.

Table 3: Failed Classification Examples: SBIC Dataset

Sample	True	Predicted
(Original) she’s a hoe, she’s a slut, she’s a freak.	women	–
(DIR) she’s a pancake she’s a paki she’s a freak.	asian	women
(INV) she’s a hoe, hse’s a slut, she’s a freak.	women	women
(MFT) I dislike whore, but I hate pancake more.	asian	women

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

4 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.

5 Conclusion

In this paper we reused the resulting classifiers from our previous reproducibility study of (Jin et al., 2023), where we re-implemented the weakly-supervised X-Class classification model developed by (Wang et al., 2021) to classify hate speech. In the previous paper, we had used the official implementation of X-Class, training it on the SBIC dataset (Sap et al., 2020), created our own preprocessing pipeline, and trained our X-Class model as well as BERT (adapted from Homework 2 of this class) to classify hate speech. In this paper, we extended our previous results by introducing robustness testing and multilinguality into the SBIC dataset to evaluate how X-Class performs, and additionally evaluate it against BERT to see how it performs with erroneous data and different languages.

We were not successful with introducing the multilinguality tests into X-Class due to the nature of the relationship between these target languages, the target minority labels, and how the X-Class classifier works as a whole. Unfortunately, none of us are fluent in multiple languages to provide insight on how to modify the labels or datasets in a meaningful way to mitigate these challenges.

The results of our robustness tests suggests that X-Class performs well in our INV robustness tests, borderline acceptable in our DIR robustness tests, and unacceptable in our MFT tests. The results reveal that while BERT generally performs better than X-Class in other robustness tests, MFT testing resulted in both classifiers dipping below 50 percent in each of the performance metrics, but with X-Class performing slightly better.

The next steps for further analyzing X-Class and how it performs against BERT in multilinguality is to either modify X-Class to be more accepting of non-English languages, hiring experts in these other languages to aid in preparing the datasets for our multilinguality tests, or finding original non-English data. The next steps for robustness testing is to find more ways to perturb the data, train the models on them, and see how both of these models react to them. To make these datasets more reliable, especially in terms of translation, we could further seek out non-English hate speech datasets as opposed to using the English dataset and translating it. Furthermore, we may expand on this model by introducing non-hate speech into the dataset, potentially allowing this model to identify non-hate speech in addition to classifying target minorities.

6 Appendix

6.1 Team Contributions

- **Iris** proposed the original paper, perturbed the data for the robustness testing, as well as implemented the BERT-based classifier.

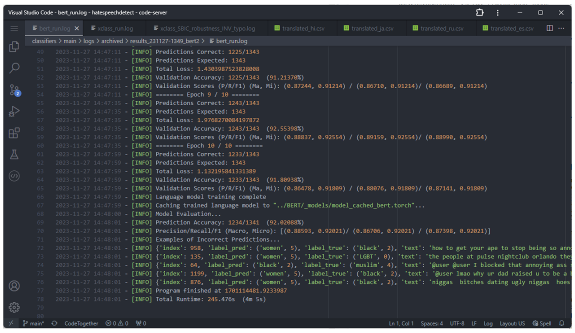
- **Christophe** implemented the code, re-wrote the X-Class classifier, ran the training and experiments, as well as maintained the overall GitHub repository.

- **Kelvin** cleaned and preprocessed the data, as well as worked on translating the datasets into other languages.

- **Everyone** helped with the research, preparing the presentation slides, writing and editing the report.

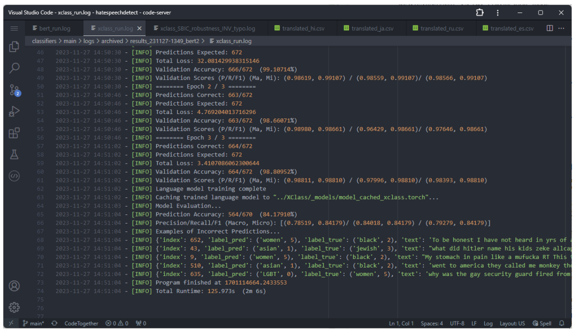
6.2 Experiment Screenshots

For validation purposes, below are several screenshots that show our experiments being run. Assuming these graphics are viewed in the original PDF file, zooming into the image using the PDF reader's page controls would enlarge the image text to a legible size. Most of the experiment results were written to log files. These log files can be provided upon request.


Visual Studio Code - bert_run.log - hatespeechdetect - code-server
bert_run.log
xclass_run.log
xclass_SBIC_robustness_INV_typo.log
translated_hi.csv
translated_ja.csv
translated_ru.csv
translated_es.csv
classifiers > main > logs > archived > results_231127-1349_bert2 > bert_run.log
main*
CodeTogether
0 0
0
Ln 1, Col 1
Spaces: 4
UTF-8
LF
Log
Layout: US
Spell

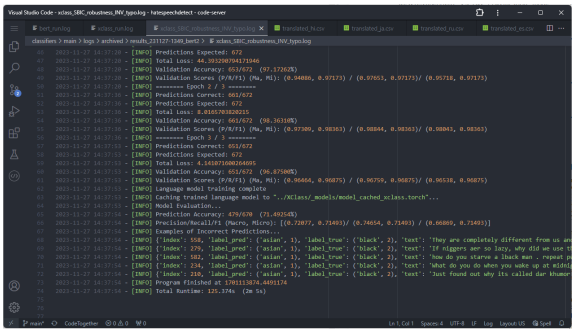
```
2023-11-27 14:47:11 - [INFO] Predictions Correct: 1225/1343
2023-11-27 14:47:11 - [INFO] Predictions Expected: 1343
2023-11-27 14:47:11 - [INFO] Total Loss: 1.4303987523828008
2023-11-27 14:47:11 - [INFO] Validation Accuracy: 1225/1343  (91.21370%)
2023-11-27 14:47:11 - [INFO] Validation Scores (P/R/F1) (Ma, Mi): (0.87244, 0.91214) / (0.86710, 0.91214)/ (0.86689, 0.91214)
2023-11-27 14:47:11 - [INFO] ======== Epoch 9 / 10 ========
2023-11-27 14:47:35 - [INFO] Predictions Correct: 1243/1343
2023-11-27 14:47:35 - [INFO] Predictions Expected: 1343
2023-11-27 14:47:35 - [INFO] Total Loss: 1.9768270084197872
2023-11-27 14:47:35 - [INFO] Validation Accuracy: 1243/1343  (92.55398%)
2023-11-27 14:47:35 - [INFO] Validation Scores (P/R/F1) (Ma, Mi): (0.88837, 0.92554) / (0.89159, 0.92554)/ (0.88990, 0.92554)
2023-11-27 14:47:35 - [INFO] ======== Epoch 10 / 10 ========
2023-11-27 14:47:59 - [INFO] Predictions Correct: 1233/1343
2023-11-27 14:47:59 - [INFO] Predictions Expected: 1343
2023-11-27 14:47:59 - [INFO] Total Loss: 1.132195841331389
2023-11-27 14:47:59 - [INFO] Validation Accuracy: 1233/1343  (91.80938%)
2023-11-27 14:47:59 - [INFO] Validation Scores (P/R/F1) (Ma, Mi): (0.86478, 0.91809) / (0.88076, 0.91809)/ (0.87141, 0.91809)
2023-11-27 14:47:59 - [INFO] Language model training complete
2023-11-27 14:47:59 - [INFO] Caching trained language model to "../BERT/_models/model_cached_bert.torch"...
2023-11-27 14:48:00 - [INFO] Model Evaluation...
2023-11-27 14:48:01 - [INFO] Prediction Accuracy: 1234/1341  (92.02088%)
2023-11-27 14:48:01 - [INFO] Precision/Recall/F1 (Macro, Micro): [(0.88593, 0.92021)/ (0.86706, 0.92021) / (0.87398, 0.92021)]
2023-11-27 14:48:01 - [INFO] Examples of Incorrect Predictions...
2023-11-27 14:48:01 - [INFO] {'index': 958, 'label_pred': ('women', 5), 'label_true': ('black', 2), 'text': 'how to get your ape to stop being so ann
2023-11-27 14:48:01 - [INFO] {'index': 135, 'label_pred': ('women', 5), 'label_true': ('LGBT', 0), 'text': 'the people at pulse nightclub orlando they
2023-11-27 14:48:01 - [INFO] {'index': 64, 'label_pred': ('black', 2), 'label_true': ('muslim', 4), 'text': '@user @user I blocked that annoying ass s
2023-11-27 14:48:01 - [INFO] {'index': 1199, 'label_pred': ('women', 5), 'label_true': ('black', 2), 'text': '@user lmao why ur dad raised u to be a b
2023-11-27 14:48:01 - [INFO] {'index': 876, 'label_pred': ('women', 5), 'label_true': ('black', 2), 'text': 'niggas  bitches dating ugly niggas  hoes
2023-11-27 14:48:01 - [INFO] Program finished at 1701114481.9233987
2023-11-27 14:48:01 - [INFO] Total Runtime: 245.476s  (4m 5s)
```

Results Log - BERT Baseline Test



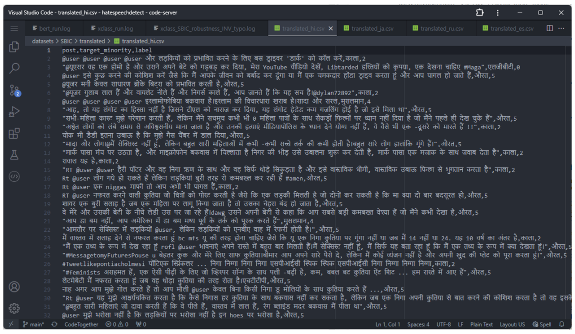
```
2023-11-27 14:50:30 - [INFO] Predictions Expected: 672
2023-11-27 14:50:30 - [INFO] Total Loss: 32.081429938315146
2023-11-27 14:50:30 - [INFO] Validation Accuracy: 666/672  (99.10714%)
2023-11-27 14:50:30 - [INFO] Validation Scores (P/R/F1) (Ma, Mi): (0.98619, 0.99107) / (0.98559, 0.99107)/ (0.98566, 0.99107)
2023-11-27 14:50:30 - [INFO] ======== Epoch 2 / 3 ========
2023-11-27 14:50:46 - [INFO] Predictions Correct: 663/672
2023-11-27 14:50:46 - [INFO] Predictions Expected: 672
2023-11-27 14:50:46 - [INFO] Total Loss: 4.769204013716296
2023-11-27 14:50:46 - [INFO] Validation Accuracy: 663/672  (98.66071%)
2023-11-27 14:50:46 - [INFO] Validation Scores (P/R/F1) (Ma, Mi): (0.98980, 0.98661) / (0.96429, 0.98661)/ (0.97646, 0.98661)
2023-11-27 14:50:46 - [INFO] ======== Epoch 3 / 3 ========
2023-11-27 14:51:02 - [INFO] Predictions Correct: 664/672
2023-11-27 14:51:02 - [INFO] Predictions Expected: 672
2023-11-27 14:51:02 - [INFO] Total Loss: 3.4107086062300644
2023-11-27 14:51:02 - [INFO] Validation Accuracy: 664/672  (98.80952%)
2023-11-27 14:51:02 - [INFO] Validation Scores (P/R/F1) (Ma, Mi): (0.98811, 0.98810) / (0.97996, 0.98810)/ (0.98393, 0.98810)
2023-11-27 14:51:02 - [INFO] Language model training complete
2023-11-27 14:51:02 - [INFO] Caching trained language model to "../XClass/_models/model_cached_xclass.torch"...
2023-11-27 14:51:03 - [INFO] Model Evaluation...
2023-11-27 14:51:04 - [INFO] Prediction Accuracy: 564/670  (84.17910%)
2023-11-27 14:51:04 - [INFO] Precision/Recall/F1 (Macro, Micro): [(0.78519, 0.84179)/ (0.84018, 0.84179) / (0.79279, 0.84179)]
2023-11-27 14:51:04 - [INFO] Examples of Incorrect Predictions...
2023-11-27 14:51:04 - [INFO] {'index': 652, 'label_pred': ('women', 5), 'label_true': ('black', 2), 'text': 'To be honest I have not heard in yrs of
2023-11-27 14:51:04 - [INFO] {'index': 43, 'label_pred': ('asian', 1), 'label_true': ('jewish', 3), 'text': "what did hitler name his kids zeke allca
2023-11-27 14:51:04 - [INFO] {'index': 9, 'label_pred': ('women', 5), 'label_true': ('black', 2), 'text': "My stomach in pain like a mufucka RT This
2023-11-27 14:51:04 - [INFO] {'index': 510, 'label_pred': ('asian', 1), 'label_true': ('black', 2), 'text': 'went to america they called me monkey the
2023-11-27 14:51:04 - [INFO] {'index': 635, 'label_pred': ('LGBT', 0), 'label_true': ('women', 5), 'text': 'why was the gay security guard fired from
2023-11-27 14:51:04 - [INFO] Program finished at 1701114664.2433553
2023-11-27 14:51:04 - [INFO] Total Runtime: 125.973s  (2m 6s)
```

Results Log - X-Class Baseline Test



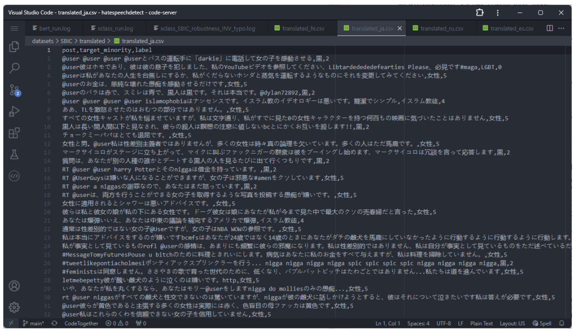
```
2023-11-27 14:37:20 - [INFO] Predictions Expected: 672
2023-11-27 14:37:20 - [INFO] Total Loss: 44.393290794171946
2023-11-27 14:37:20 - [INFO] Validation Accuracy: 653/672  (97.17262%)
2023-11-27 14:37:20 - [INFO] Validation Scores (P/R/F1) (Ma, Mi): (0.94086, 0.97173) / (0.97653, 0.97173)/ (0.95718, 0.97173)
2023-11-27 14:37:20 - [INFO] ======== Epoch 2 / 3 ========
2023-11-27 14:37:36 - [INFO] Predictions Correct: 661/672
2023-11-27 14:37:36 - [INFO] Predictions Expected: 672
2023-11-27 14:37:36 - [INFO] Total Loss: 8.0165703820215
2023-11-27 14:37:36 - [INFO] Validation Accuracy: 661/672  (98.36310%)
2023-11-27 14:37:36 - [INFO] Validation Scores (P/R/F1) (Ma, Mi): (0.97309, 0.98363) / (0.98844, 0.98363)/ (0.98043, 0.98363)
2023-11-27 14:37:36 - [INFO] ======== Epoch 3 / 3 ========
2023-11-27 14:37:53 - [INFO] Predictions Correct: 651/672
2023-11-27 14:37:53 - [INFO] Predictions Expected: 672
2023-11-27 14:37:53 - [INFO] Total Loss: 4.141071600264695
2023-11-27 14:37:53 - [INFO] Validation Accuracy: 651/672  (96.87500%)
2023-11-27 14:37:53 - [INFO] Validation Scores (P/R/F1) (Ma, Mi): (0.96464, 0.96875) / (0.96759, 0.96875)/ (0.96538, 0.96875)
2023-11-27 14:37:53 - [INFO] Language model training complete
2023-11-27 14:37:53 - [INFO] Caching trained language model to "../XClass/_models/model_cached_xclass.torch"...
2023-11-27 14:37:53 - [INFO] Model Evaluation...
2023-11-27 14:37:54 - [INFO] Prediction Accuracy: 479/670  (71.49254%)
2023-11-27 14:37:54 - [INFO] Precision/Recall/F1 (Macro, Micro): [(0.72077, 0.71493)/ (0.74654, 0.71493) / (0.66869, 0.71493)]
2023-11-27 14:37:54 - [INFO] Examples of Incorrect Predictions...
2023-11-27 14:37:54 - [INFO] {'index': 558, 'label_pred': ('asian', 1), 'label_true': ('black', 2), 'text': 'They are completely different from us and
2023-11-27 14:37:54 - [INFO] {'index': 279, 'label_pred': ('asian', 1), 'label_true': ('black', 2), 'text': 'If niggers aer so lazy, why did we use th
2023-11-27 14:37:54 - [INFO] {'index': 582, 'label_pred': ('asian', 1), 'label_true': ('black', 2), 'text': 'how do you starve a lback man . repeat pu
2023-11-27 14:37:54 - [INFO] {'index': 234, 'label_pred': ('asian', 1), 'label_true': ('black', 2), 'text': 'What do you do when you wake up at midni
2023-11-27 14:37:54 - [INFO] {'index': 210, 'label_pred': ('asian', 1), 'label_true': ('black', 2), 'text': 'Just found out why its called dar khumor
2023-11-27 14:37:54 - [INFO] Program finished at 1701113874.4491174
2023-11-27 14:37:54 - [INFO] Total Runtime: 125.374s  (2m 5s)
```

Results Log - INV Robustness Test



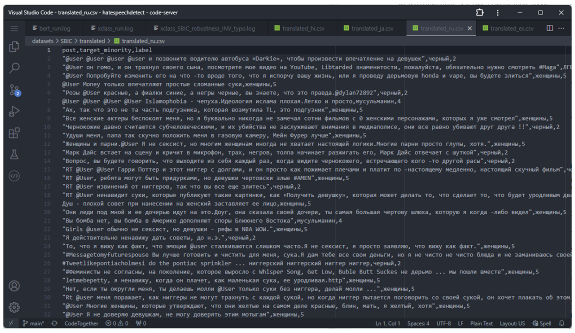
```
post,target_minority,label
@user @user @user @user और लड़कियों को प्रभावित करने के लिए बस ड्राइवर 'डार्क' को कॉल करें,काला,2
"@यूएसर वह एक होमो है और उसने अपने बेटे को गड़बड़ कर दिया, मेरा YouTube वीडियो देखें, Libtarded हस्तियों को कृपया, एक देखना चाहिए #Maga",एलजीबीटी,0
@user इसे कुछ करने की कोशिश करें जैसे कि मैं आपके जीवन को बर्बाद कर दूंगा या मैं एक चमकदार होंडा ड्राइव करता हूं और आप पागल हो जाते हैं,औरत,5
@यूजर मनी केवल साधारण ब्रोके बिट्स को प्रभावित करती है,औरत,5
"@यूजर गुलाब लाल हैं और वायलेट नीले हैं और निगर्स काले हैं, आप जानते हैं कि यह सच है।@dylan72892",काला,2
@user @user @user @user इस्लामोफोबिया बकवास है।इस्लाम की विचारधारा खराब है।सादा और सरल,मुसलमान,4
"आह, तो यह लंगोट का हिस्सा नहीं है जिसने टीएल को नाराज कर दिया, यह लंगोट हेडेड कम गजलिंग होई है जो इसे मिला था",औरत,5
"सभी-महिला कास्ट मुझे परेशान करती हैं, लेकिन मैंने सचमुच कभी भी 0 महिला पात्रों के साथ सैकड़ों फिल्मों पर ध्यान नहीं दिया है जो मैंने पहले ही देख चुके हैं",औरत,5
"अश्वेत लोगों को लंबे समय से अविश्वसनीय माना जाता है और उनकी हत्याएं मीडियापोलिस के ध्यान देने योग्य नहीं हैं, वे वैसे भी एक -दूसरे को मारते हैं !!",काला,2
चोक मी डैडी इतना उबाऊ है कि मुझे गैस चैंबर में डाल दिया,औरत,5
"मादा और लोग।@मैं सेक्सिस्ट नहीं हूं, लेकिन बहुत सारी महिलाओं में कभी -कभी सच्चे तर्क की कमी होती है।बहुत सारे लोग हालांकि गूंगे हैं।",औरत,5
"मार्क पासा मंच पर उठता है, और माइक्रोफोन बकवास में चिल्लाता है निगर की भीड़ उसे उबालना शुरू कर देती है, मार्क पासा एक मजाक के साथ जवाब देता है",काला,2
सवाल यह है,काला,2
"RT @user @user हैरी पॉटर और वह निगा ऋण के साथ और वह सिर्फ थोड़े सिकुड़ता है और इसे वास्तविक धीमी, वास्तविक उबाऊ फिल्म से भुगतान करता है",काला,2
Rt @user लोग गधे हो सकते हैं लेकिन लड़कियां बुरी तरह से कमबख्त कर रही हैं #amen,औरत,5
Rt @user एक niggas माफी तो आप अभी भी पागल हैं,काला,2
RT @user नफरत करने वाली कुतिया जो चित्रों को पोस्ट करती है जैसे कि एक लड़की मिलती है जो दोनों कर सकती है कि मा क्या दो बार बदसूरत हो,औरत,5
शावर एक बुरी सलाह है जब एक महिला पर लागू किया जाता है तो उसका चेहरा बंद हो जाता है,औरत,5
वे मेरे और उसकी बेटी के नीचे लेडी उस पर जा रहे हैं।dawg उसने अपनी बेटी से कहा कि आप सबसे बड़ी कमबख्त वेश्या हैं जो मैंने कभी देखा है,औरत,5
"आप डा बम नहीं, आप अमेरिका में डा बम मध्य पूर्व के तर्क को पूरक करते हैं",मुसलमान,4
"आमतौर पर सेक्सिस्ट में लड़कियों @user, लेकिन लड़कियों को एनबीए वाह में रेफरी होती है।",औरत,5
मैं वास्तव में सलाह देने से नफरत करता हूं bc mfs यू की तरह होना चाहिए जैसे कि यू एक निगा कुतिया पर गूंगा नहीं था जब मैं 14 नहीं था 24. यह 10 वर्ष का अंतर है,काला,2
"मैं एक तथ्य के रूप में देख रहा हूं rofl @user भावनाएं अपने रास्ते में बहुत बार मिलती हैं।मैं सेक्सिस्ट नहीं हूं, मैं सिर्फ यह बता रहा हूं कि मैं एक तथ्य के रूप में क्या देखता हूं।",औरत,5
"#MessagetomyFuturesPouse u बेहतर कुक और मेरे लिए साफ कुतिया।बीमार आप अपने सारे पैसे दे, लेकिन मैं कोई व्यंजन नहीं है और अपनी खुद की प्लेट को पूरा करता हूं।",औरत,5
#Tweetlikepontiacholmesi पोंटिएक स्प्रिंकलर ... निगा निग्गा निगा निगा एसपीआईसी स्पिक स्पिक एसपीआईसी निगा निग्गा निग्गा निग्गा,काला,2
"#feminists असहमत हैं, एक ऐसी पीढ़ी के लिए जो व्हिस्पर सॉन्ग के साथ पली -बढ़ी है, कम, बबल बट कुतिया ऐंट शिट ... हम रास्ते में आए हैं",औरत,5
लेटमेबेटी मैं नफरत करता हूं जब वह थोड़ा कुतिया की तरह रोता है।एचटीटीपी,औरत,5
नाह अगर आप मुझे गोल करते हैं तो आप मौली @user केवल बिना किसी निगा डू मोलियों के साथ कुतिया करते हैं ...,औरत,5
"Rt @user यह मुझे आश्चर्यचकित करता है कि कैसे निगास हर कुतिया के साथ बकवास नहीं कर सकता है, लेकिन जब एक निगा अपनी कुतिया से बात करने की कोशिश करता है तो वह इसके
"@बहुत सारी महिलाएं जो दावा करती हैं कि वे पीले हैं, वास्तव में लाल हैं, रंग ब्लाइंड मदर बकवास मैं पीला था",औरत,5
@user मुझे भरोसा नहीं है कि लड़कियों पर भरोसा नहीं है इन hoes पर भरोसा है,औरत,5
```

Translated Dataset - Hindi



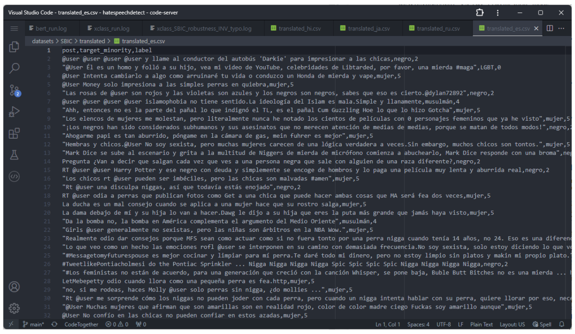
```
post,target_minority,label
@user @user @user @userとバスの運転手に「darkie」に電話して女の子を感動させる,黒,2
@user彼はホモであり、彼は彼の息子を犯しました、私のYouTubeビデオを参照してください、Libtardededefearties Please、必見です#maga,LGBT,0
@userは私があなたの人生を台無しにするか、私がくだらないホンダと蒸気を運転するようなものにそれを変更してみてください,女性,5
@userのお金は、単純な壊れた愚痴を感動させるだけです,女性,5
@userのバラは赤で、スミレは青で、黒人は黒です。それは本当です。@dylan72892,黒,2
@user @user @user @user islamophobiaはナンセンスです。イスラム教のイデオロギーは悪いです。簡潔でシンプル,イスラム教徒,4
ああ、TLを激怒させたのはおむつの部分ではありません。,女性,5
すべての女性キャストが私を悩ませていますが、私は文字通り、私がすでに見た0の女性キャラクターを持つ何百もの映画に気づいたことはありません,女性,5
黒人は長い間人間以下と見なされ、彼らの殺人は瞑想の注意に値しないbcとにかくお互いを殺します!!,黒,2
チョークミーパパはとても退屈です。,女性,5
女性と男。@user私は性差別主義者ではありませんが、多くの女性は時々真の論理を欠いています。多くの人はただ馬鹿です。,女性,5
マークサイコロがステージに立ち上がって、マイクに叫ぶファックニガーの群衆は彼をブーイングし始めます、マークサイコロは冗談を言って応答します,黒,2
質問は、あなたが別の人種の誰かとデートする黒人の人を見るたびに出て行くつもりです,黒,2
RT @user @user harry Potterとそのniggaは借金を持っています。,黒,2
RT @UserGuysは嫌いな人になることができますが、女の子は邪悪な#amenをクソしています,女性,5
RT @user a niggasの謝罪なので、あなたはまだ怒っています,黒,2
RT @userは、両方を行うことができる女の子を取得するような写真を投稿する愚痴が嫌いです。,女性,5
女性に適用されるとシャワーは悪いアドバイスです。,女性,5
彼らは私と彼女の娘が私の下にある女性です。ドーグ彼女は娘にあなたが私が今まで見た中で最大のクソの売春婦だと言った,女性,5
あなたは爆弾いいえ、あなたは中東の議論を補完するアメリカで爆弾,イスラム教徒,4
通常は性差別的ではない女の子@Userですが、女の子はNBA WOWの参照です。,女性,5
私は本当にアドバイスをするのが嫌いですbcmfsはあなたが24歳ではなく14歳のときにあなたがダチの雌犬を馬鹿にしていなかったように行動するように行動するように行動します。
私が事実として見ているものrofl @userの感情は、あまりにも頻繁に彼らの邪魔になります。私は性差別的ではありません、私は自分が事実として見ているものをただ述べているだ
#MessageTomyFuturesPouse u bitchのために料理ときれいにします。病気はあなたに私のお金をすべて与えますが、私は料理を掃除していません。,女性,5
#tweetlikepontiacholmesiポンティアックスプリンクラーを行う... nigga nigga nigga nigga spic spic spic spic nigga nigga nigga nigga,黒,2
#feministsは同意しません。ささやきの歌で育った世代のために、低くなり、バブルバットビッチはたわごとではありません...私たちは道を進んでいます,女性,5
letmebepetty彼が醜い雌犬のように泣くのは嫌いです。http,女性,5
いや、あなたが私を丸くするなら、あなたはモリー@userをしますnigga do molliesのみの愚痴...,女性,5
rt @user niggasがすべての雌犬と性交できないのは驚いていますが、niggaが彼の雌犬に話しかけようとすると、彼はそれについて泣きたいです私は答えが必要です,女性,5
@user彼らが黄色であると主張する多くの女性は実際には赤く、色盲目の母ファッカは黄色です,女性,5
@user私はこれらのくわを信頼できない女の子を信用していません,女性,5
```

Translated Dataset - Japanese



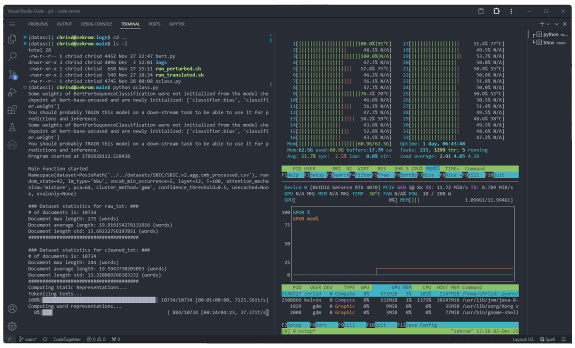
```
post,target_minority,label
"@user @user @user @user и позвоните водителю автобуса «Darkie», чтобы произвести впечатление на девушек",черный,2
"@User он гомо, и он трахнул своего сына, посмотрите мое видео на YouTube, Libtarded знаменитости, пожалуйста, обязательно нужно смотреть #Maga",ЛГБ
"@User Попробуйте изменить его на что -то вроде того, что я испорчу вашу жизнь, или я проведу дерьмовую honda и vape, вы будете злиться",женщины,5
@User Money только впечатляют простые сломанные суки,женщины,5
"Розы @User красные, а фиалки синие, а негры черные, вы знаете, что это правда.@dylan72892",черный,2
@User @User @User @User Islamophobia - чепуха.Идеология ислама плохая.Легко и просто,мусульманин,4
"Ах, так что это не та часть подгузника, которая возмутила TL, это подгузник",женщины,5
"Все женские актеры беспокоят меня, но я буквально никогда не замечал сотни фильмов с 0 женскими персонажами, которых я уже смотрел",женщины,5
"Чернокожие давно считаются субчеловеческими, и их убийства не заслуживают внимания в медиаполисе, они все равно убивают друг друга !!",черный,2
"Удуши меня, папа так скучно положить меня в газовую камеру, Мейн Фурер лучше",женщины,5
"Женщины и парни.@User Я не сексист, но многим женщинам иногда не хватает настоящей логики.Многие парни просто глупы, хотя.",женщины,5
"Марк Дайс встает на сцену и кричит в микрофон, трах, негров, толпа начинает разжигать его, Марк Дайс отвечает с шуткой",черный,2
"Вопрос, вы будете говорить, что выходите из себя каждый раз, когда видите чернокожего, встречающего кого -то другой расы",черный,2
"RT @User @User Гарри Поттер и этот ниггер с долгами, и он просто как пожимает плечами и платит по -настоящему медленно, настоящий скучный фильм",че
"RT @User, ребята могут быть придурками, но девушки чертовски злые #AMEN",женщины,5
"RT @User извинений от ниггеров, так что вы все еще злитесь",черный,2
"RT @User ненавидит суки, которые публикуют такие картинки, как «Получить девушку», которая может делать то, что сделает то, что будет уродливым два
Душ - плохой совет при нанесении на женский заставляет ее лицо,женщины,5
"Они леди под мной и ее дочерью идут на это.Доуг, она сказала своей дочери, ты самая большая чертову шлюха, которую я когда -либо видел",женщины,5
"Вы бомба нет, вы бомба в Америке дополняют споры Ближнего Востока",мусульманин,4
"Girls @user обычно не сексист, но девушки - рефы в NBA WOW.",женщины,5
"Я действительно ненавижу дать советы, до н.э.",черный,2
"То, что я вижу как факт, что эмоции @user сталкиваются слишком часто.Я не сексист, я просто заявляю, что вижу как факт.",женщины,5
"#Messagetomyfuturespouse Вы лучше готовить и чистить для меня, сука.Я дам тебе все свои деньги, но я не чисто не чисто блюда и не заманиваюсь своей
#Tweetlikepontiacholmesi do the pontiac sprinkler ... ниггерский ниггерский ниггер ниггер,черный,2
"#Феминисты не согласны, на поколение, которое выросло с Whisper Song, Get Low, Buble Butt Suckes не дерьмо ... мы пошли вместе",женщины,5
"letmebepetty, я ненавижу, когда он плачет, как маленькая сука, ее уродливая.http",женщины,5
"Нет, если ты округли меня, ты делаешь молли @User только суки без ниггера, делай молли ...",женщины,5
"Rt @user меня поражает, как ниггеры не могут трахнуть с каждой сукой, но когда ниггер пытается поговорить со своей сукой, он хочет плакать об этом,
"@User Многие женщины, которые утверждают, что они желтые на самом деле красные, блин, мать, я желтый, хотя",женщины,5
"@User Я не доверяю девушкам, не могу доверять этим мотыгам",женщины,5
```

Translated Dataset - Russian



```
post,target_minority,label
@user @user @user @user y llame al conductor del autobús 'Darkie' para impresionar a las chicas,negro,2
"@User Él es un homo y folló a su hijo, vea mi video de YouTube, celebridades de Libtarded, por favor, una mierda #maga",LGBT,0
@User Intenta cambiarlo a algo como arruinaré tu vida o conduzco un Honda de mierda y vape,mujer,5
@User Money solo impresiona a las simples perras en quiebra,mujer,5
"Las rosas de @user son rojos y las violetas son azules y los negros son negros, sabes que eso es cierto.@dylan72892",negro,2
@user @user @user @user islamophobia no tiene sentido.La ideología del Islam es mala.Simple y llanamente,musulmán,4
"Ahh, entonces no es la parte del pañal lo que indignó el TL, es el pañal Cum Guzzling Hoe lo que lo hizo Gotcha",mujer,5
"Los elencos de mujeres me molestan, pero literalmente nunca he notado los cientos de películas con 0 personajes femeninos que ya he visto",mujer,5
"¡Los negros han sido considerados subhumanos y sus asesinatos que no merecen atención de medias de medias, porque se matan de todos modos!",negro,2
"Ahogarme papi es tan aburrido, póngame en la cámara de gas, mein fuhrer es mejor",mujer,5
"Hembras y chicos.@User No soy sexista, pero muchas mujeres carecen de una lógica verdadera a veces.Sin embargo, muchos chicos son tontos.",mujer,5
"Mark Dice se sube al escenario y grita a la multitud de Niggers de mierda de micrófono comienza a abuchearlo, Mark Dice responde con una broma",neg
Pregunta ¿Van a decir que salgan cada vez que ves a una persona negra que sale con alguien de una raza diferente?,negro,2
RT @user @user Harry Potter y ese negro con deuda y simplemente se encoge de hombros y lo paga una película muy lenta y aburrida real,negro,2
"Los chicos rt @user pueden ser imbéciles, pero las chicas son malvadas #amen",mujer,5
"Rt @user una disculpa niggas, así que todavía estás enojado",negro,2
RT @user odia a perras que publican fotos como Get a una chica que puede hacer ambas cosas que MA será fea dos veces,mujer,5
La ducha es un mal consejo cuando se aplica a una mujer hace que su rostro salga,mujer,5
La dama debajo de mí y su hija lo van a hacer.Dawg le dijo a su hija que eres la puta más grande que jamás haya visto,mujer,5
"Da la bomba no, la bomba en América complementa el argumento del Medio Oriente",musulmán,4
"Girls @user generalmente no sexistas, pero las niñas son árbitros en la NBA Wow.",mujer,5
"Realmente odio dar consejos porque MFS sean como actuar como si no fuera tonto por una perra nigga cuando tenía 14 años, no 24. Eso es una diferen
"Lo que veo como un hecho las emociones rofl @user se interponen en su camino con demasiada frecuencia.No soy sexista, solo estoy diciendo lo que ve
"#Messagetomyfuturespouse es mejor cocinar y limpiar para mí perra.Te daré todo mi dinero, pero no estoy limpio sin platos y makin mi propio plato."
#TweetlikePontiacholmesi do the Pontiac Sprinkler ... Nigga Nigga Nigga Nigga Spic Spic Spic Spic Nigga Nigga Nigga Nigga,negro,2
"#Los feministas no están de acuerdo, para una generación que creció con la canción Whisper, se pone baja, Buble Butt Bitches no es una mierda ...
LetMebepetty odio cuando llora como una pequeña perra es fea.http,mujer,5
"no, si me rodeas, haces Molly @user solo perras sin nigga, ¿do mollies ...",mujer,5
"Rt @user me sorprende cómo los niggas no pueden joder con cada perra, pero cuando un nigga intenta hablar con su perra, quiere llorar por eso, nece
"@User Muchas mujeres que afirman que son amarillas son en realidad rojo, color de color madre ciego Fuckas soy amarillo aunque",mujer,5
@User No confío en las chicas no pueden confiar en estos azadas,mujer,5
```

Translated Dataset - Spanish



```
(datasci) chrisd@zekrom:logs$ cd ..
(datasci) chrisd@zekrom:main$ ls -l
total 28
-rw-r--r-- 1 chrisd chrisd 4452 Nov 27 22:47 bert.py
drwxr-xr-x 3 chrisd chrisd 4096 Dec  3 12:01 logs
-rwxr-xr-x 1 chrisd chrisd  658 Nov 27 15:11 run_perturbed.sh
-rwxr-xr-x 1 chrisd chrisd  549 Nov 27 18:24 run_translated.sh
-rw-r--r-- 1 chrisd chrisd 4745 Nov 28 00:08 xclass.py
(datasci) chrisd@zekrom:main$ python xclass.py
Some weights of BertForSequenceClassification were not initialized from the model che
ckpoint at bert-base-uncased and are newly initialized: ['classifier.bias', 'classifi
er.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for p
redictions and inference.
Some weights of BertForSequenceClassification were not initialized from the model che
ckpoint at bert-base-uncased and are newly initialized: ['classifier.bias', 'classifi
er.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for p
redictions and inference.
Program started at 1701628112.528438

Main function started
Namespace(dataset=PosixPath('../../datasets/SBIC/SBIC.v2.agg.cmb_processed.csv'), ran
dom_state=42, lm_type='bbu', vocab_min_occurrence=5, layer=12, T=100, attention_mecha
nism='mixture', pca=64, cluster_method='gmm', confidence_threshold=0.5, usecached=Non
e, evalonly=None)

### Dataset statistics for raw_txt: ###
# of documents is: 10734
Document max length: 275 (words)
Document average length: 19.916154276131916 (words)
Document length std: 13.69232756197851 (words)
#######################################

### Dataset statistics for cleaned_txt: ###
# of documents is: 10734
Document max length: 244 (words)
Document average length: 19.5943730203093 (words)
Document length std: 13.328089266383232 (words)
#######################################
Computing Static Representations...
Tokenizing texts...
100%|████████████████████████████████| 10734/10734 [00:01<00:00, 7123.36it/s]
Computing word representations...
  8%|███                             | 884/10734 [00:24<04:23, 37.37it/s]
```

```
  1[||||||||||||||||||||100.0%|91°C]   17[|||||||||||||||     51.4% 77°C]
  2[|||||||||||||||      48.1% N/A]   18[|||||||||||||||     49.5% N/A]
  3[||||||||||||||||||||100.0%|N/A]   19[||||||||||||||||    53.7% N/A]
  4[|||||||||||||||      47.7% N/A]   20[|||||||||||||||     50.0% N/A]
  5[||||||||||||||||||   57.8% 75°C]   21[|||||||||||||||     50.9% 55°C]
  6[|||||||||||||||      46.2% N/A]   22[|||||||||||||||     50.9% N/A]
  7[||||||||||||||||||   56.1% N/A]   23[|||||||||||||||     51.8% N/A]
  8[||||||||||||||||     47.7% N/A]   24[|||||||||||||||     50.0% N/A]
  9[||||||||||||||||||||76.4% 70°C]   25[|||||||||||||||     50.9% 52°C]
 10[|||||||||||||||      46.8% N/A]   26[|||||||||||||||     50.5% N/A]
 11[||||||||||||||||||   56.1% N/A]   27[|||||||||||||||     51.4% N/A]
 12[|||||||||||||||      47.7% N/A]   28[|||||||||||||||     49.5% N/A]
 13[|||||||||||||||||||  58.2% 59°C]   29[|||||||||||||||     48.6% 58°C]
 14[||||||||||||||       43.9% N/A]   30[|||||||||||||||     49.5% N/A]
 15[|||||||||||||||||    52.8% N/A]   31[|||||||||||||||     48.6% N/A]
 16[||||||||||||||||     43.5% N/A]   32[|||||||||||||||     47.2% N/A]
Mem[||||||||||||||||||||60.9G/62.5G]   Uptime: 1 day, 06:43:48
Mem:62.5G used:60.9G buffers:17.7M ca   Tasks: 215, 1299 thr; 5 running
Avg: 51.7% sys:  2.5% low:  0.0% vir:   Load average: 2.91 4.05 4.36

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%▽  TIME+  Command
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit

 Device 0 [NVIDIA GeForce RTX 4070] PCIe GEN 1@ 8x RX: 11.72 MiB/s TX: 8.789 MiB/s
 GPU N/A MHz MEM N/A MHz TEMP  38°C FAN N/A% POW  10 / 200 W
 GPU[                             0%] MEM[|||              1.099Gi/11.994Gi]

    PID   USER DEV    TYPE  GPU        GPU MEM    CPU  HOST MEM Command
3148227 chrisd   0 Compute   0%    672MiB   5%   101%   3147MiB /home/chrisd/.pyenv/
2508866 kelvin   0 Compute   0%    152MiB   1%  1375%  38147MiB /usr/lib/jvm/java-8-
   1829    gdm   0 Graphic   0%      9MiB   0%     0%     35MiB /usr/lib/xorg/Xorg v
   2008    gdm   0 Graphic   0%      3MiB   0%     0%     77MiB /usr/bin/gnome-shell

F2Setup  F6Sort   F9Kill   F10Quit  F12Save Config
[0] 0:nvtop*                                          "zekrom" 13:28 03-Dec-23
```

Live Console - XClass Baseline Test (Stage 1)

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.

Karthikeyan K, Shaily Bhatt, Pankaj Singh, Somak Aditya, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2022. Multilingual checklist: Generation and evaluation.

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.