

Testing the Viability of Synthetic Datasets for the Hate Speech Detection Task: A Qualitative Study

Ian O. Rodríguez Almodóvar
rodriguezio@cardiff.ac.uk

Omar Aqel Ibrahim
ibrahimo@cardiff.ac.uk

Abstract

With the influx of large language models (LLMs), hate speech detection has seen substantial advances, especially through their ability to augment textual data. In this paper, we aim to synthesize the training set of the Measuring Hate Speech Dataset (Kennedy et al., 2020) using three decoders—Mistral-7B-Instruct-V1 (Jiang et al., 2023), Mixtral 7x8B-Instruct-V1 (Jiang et al., 2024), and Falcon-7B (Almazrouei et al., 2023). We then use two encoders—RoBERTa-Base & RoBERTa-Large to classify the gold and synthetic datasets. Our results show that Falcon synthesis helped in improving generalization by achieving performance comparable to gold-only models, and even generalize better in some out-of-distribution settings. However, the corpora generated by Mistral and Mixtral is deemed to be linguistically unrealistic due to the possessed vulgarity filter in the decoders.

Warning

This paper contains examples that may be offensive or upsetting.

1 Introduction

This paper intends to replicate and expand upon a paper titled *Delving into Qualitative Implications of Synthetic Data for Hate Speech Detection* by Camila Casula et al. (2024). In this paper, the scientists sought to investigate how useful synthetically generated data is for the task of Hate Speech Detection. Methods for the Hate Speech Detection task have been widely investigated in the field of computational linguistics, and they tend to rely on gold label, annotated datasets; some of which are extensive in scope, but have their limitations. The possibility of using synthetic data for this

task is an uncharted territory, and we wanted to expand said study by observing how other models (especially ones that do not have a vulgarity filter) might perform. The main idea for the expansion was to find a model that could generate synthetic data that emulates real samples as realistically as possible. This required finding a model that did not have a vulgarity filter. The chosen model was Falcon-7B, a decoder only model that was trained on mostly English and French, and on a large-scale corpora of web data. Falcon-7B was successfully able to generate more realistic paraphrases of hate speech text than the two models used in the study (Mistral 7B/Mixtral 8x7B). Moreover, the aforementioned paper had proven that the synthetic data they had generated failed to represent the linguistic nuances required to reliably detect hate speech, which is a notoriously difficult task that is subjective and complex in nature. Thus, part of the rationale for choosing Falcon-7B, was that in generating realistic data, that it would capture more of that much needed linguistic nuance in order to correctly label out-of-distribution instances of hate speech. This study proves that generating synthetic data which is more realistic, lacks vulgarity filter, and a more detailed prompt (by using Falcon 7B) helps the classifiers better detect real instances of hate speech; albeit with some limitations.

2 Related Work/Background

The realm of synthesizing hate speech datasets to improve robustness and generalization has seen a spark with the emergence of LLMs. For example, (Pendzel et al., 2023) allocated 5 small hate datasets and fine-tuned GPT-2 with them. They then ask GPT-2 to generate over one million synthetic samples, which were fil-

tered using a BERT classifier for quality. This synthetic data, augmented with the original data, increased the score of recall and generalization across domains, though it slightly reduced precision. Another research that used GPT-2 as well is Wullach et al. (2020), in which the authors generated over one million synthetic hate and non-hate examples using GPT-2. Their approach involved fine-tuning separate models for each class and dataset, followed by filtering the outputs with a BERT classifier to ensure label accuracy and linguistic variety. The resulting dataset, named MegaSpeech, helped improve recall and generalization, particularly in cross-dataset settings where traditional models often struggle. However, since the task of hate speech detection is highly dependent on the context, early models like GPT-2 failed to capture contextual nuances. Therefore, (Girón et al., 2025) argued that Mistral allows users to generate hate speech discourse without the need for unethical jailbreaking. Their results show that classifiers trained on *Mistral-7B-Instruct-v0.2* outperformed traditional methods and simpler LLM prompts, especially under low-resource conditions. The advent Mistral 7B (Struppek et al., 2024) prompted *Mixtral-8x7B-Instruct* to craft hate speech variants that circumvented being identified by the BERT-based classifier. Mixtral made tiny adjustments such as (like bitch → bltch). Unlike GPT-4 or LLaMA2, Mixtral didn’t block these requests, which made it ideal for this kind of adversarial testing. The model achieved a 90% success rate in fooling the classifier, showing how easily these tools can be used to generate harmful content undetected — especially when safety filters are absent. However, the paper we are replicating (Casula et al., 2024) states that Mistral and Mixtral, although generalized on out-of-distribution datasets, could not generate linguistically realistic examples. The models struggled with generating vulgar words due to toxicity filter they possess and the simple prompt they used. Therefore, in this study we aim to deploy an additional model, Falcon 7B, to realistically synthesize The MHS dataset. Furthermore, to force creativity, we will use a prompt that forcibly generate realistic instances.

3 Experimentation

This section delineates first the configuration strategies utilized for the Mistral 7B and Mixtral 8x7B models, which form part of the replication section of this paper. The configuration strategy decided for Falcon 7B and the rationale for it is separate and part of the expansion section of this paper.

3.1 Data Preprocessing

The dataset used in this study requires three preprocessing steps to render it suitable for synthesization and fine-tuning.

The first preprocessing step is about aggregating multiple annotations for each sentence. Since each comment may have been labeled by multiple annotators, we group the data by *comment_id* and compute the average of the *hate_speech_score* values. This step paves the way for assigning hate-non-hate labels. We, subsequently aggregate the hate speech score as following; if the hate score is above 1, the sentence is deemed hateful. If below 1, it is deemed non-hateful. If it is 1.0, it is considered ambiguous and therefore removed.

The second preprocessing step is about marking the identity group columns (Race, Religion, Sexuality) in a binary approach. This is done by navigating the subgroups for each category in the dataset. If it is found that half of the annotators agreed that a comment targeted a subgroup within that category, we assigned a binary label of 1 to the identity group; otherwise, it was labeled 0.

The third preprocessing step splits the dataset the same way it is split in the paper. The split was specified to be 85% allocated for training, 10% for testing, and 5% for the development set.

It is important to note that the authors suggested the implementation of further preprocessing steps, but were kept out of the paper.

3.2 Replication: Mistral 7B & Mixtral 8x7B

The steps taken to generate synthetic data with Mistral 7B and Mixtral 8x7B are nearly identical to those delineated in the research paper that inspired this expansion: *Delving into Qualitative Implications of Synthetic Data*

for *Hate Speech Detection*. Relating to the decoder models, the Mistral 7B model was loaded with its default values, but the Mixtral 8x7B model needed 4-bit quantization in order to reduce memory usage due to resource limitations (this was done with the *bitsandbytes* library. *Top-p* decoding was kept 0.9+. When utilizing 4-bit quantization to help with model efficiency there might be a slight performance trade off (since 4-bit quantization makes model loading more quick and efficient), therefore it was decided that for Mistral 7B and Falcon 7B that we would load them at their default 32 bits. In order to be able to test and keep track of generation, it was decided to generate in three splits; each of which would take around 3-4 hours to train for each model. The prompt was identical to the one mentioned in the paper:

```
<s>[INST] Paraphrase this text: "text"
[/INST]
```

Since the specific hyperparameters used for prompting are not mentioned in the aforementioned paper, we have decided to implement our own. Observe the table below:

| Hyperparameter | Value |
|----------------|------------------------|
| max_new_tokens | 75 |
| do_sample | True |
| temperature | 0.9 |
| top_p | 0.95 |
| top_k | 50 |
| pad_token_id | tokenizer.eos_token_id |

Table 1: Hyperparameters used for prompting

As displayed, the temperature is set at 0.9 to ensure a high amount of creativity; and *max_new_tokens* is set at 75 to ensure that paraphrases generated by the models are long enough to encapsulate each instance, but not so long so as to overcharge the model itself, which would further drain resources and make generation even slower.

3.2.1 Fine-Tuning

The methodology for fine-tuning dictated an across-the-board application of the same parameters for each model instance. Each classifier model was fine-tuned a total of four times on four different datasets, corresponding to:

1. Synthetic-only, no classifier
2. Synthetic-only, with classifier

3. Synthetic-Gold combined, no classifier
4. Synthetic-Gold combined, with classifier

The no-classifier datasets used for fine-tuning were passed only through the *thefuzz* library, which removes synthetic instances of hate speech that are too similar to the gold instances; for this, we set a similarity threshold of 75%. The classifier datasets however, were passed through a RoBERTa Base or Large that was trained on the gold data, in order to discard synthetic examples for which a model has attributed a different label to the one that was originally assigned to it in the gold dataset. As per the methodology delineated in the paper, the default training parameters for the two classifiers (RoBERTa Base and RoBERTa Large) from the *TrainingArgs* class were used, namely setting the batch size to 64, maximum sequence length to 150, and the learning rate kept at 5e-6. The models were trained for 3 epochs.

3.3 Expansion: Falcon 7B

Another very important purpose of our paper was to investigate the viability of utilizing Falcon 7B to produce more reliably realistic instances of hate speech. To accomplish this, we implemented the most ideal hyperparameters possible:

| Hyperparameter | Value |
|----------------|------------------------|
| max_new_tokens | 100 |
| do_sample | True |
| temperature | 0.95 |
| top_p | 0.92 |
| top_k | 40 |
| pad_token_id | tokenizer.eos_token_id |

Table 2: Hyperparameters used for text generation.

Furthermore, we decided that the prompt used in the paper this experiment is based on was too short and too unspecific; therefore, we thought it best to experiment with a more specific instructional approach. Observe the following prompt below:

```
Paraphrase this sentence using casual,
vulgar language. Do not repeat the
sentence.
```

This prompting approach allows the model

to be creative in its synthesis while maintaining a lean level of specificity. During testing of the Falcon-generated splits, it was noticed that Falcon would simply repeat the sentence itself, hence the additional sentence of the prompt to rectify this problem.

3.3.1 Falcon 7B Fine-Tuning

When it comes to the filtering, the steps used for the Falcon-generated datasets were identical to those applied to the Mistral/Mixtral datasets. As for fine-tuning, the configurations for the classifiers (RoBERTa Base and Large) were also the exact same.

4 Results and Evaluation

The intent of the results and evaluation section of this paper is to point out the results obtained through our experiments. This section will thus be reserved to making these comparisons between the models and pointing out whatever similarities and discrepancies we have found. It is imperative to note that some of the results when it comes to the out-of-distribution performance may be different due to a missing preprocessing step in the paper that these experiments were inspired by.

4.1 Results: Mistral 7B

4.1.1 RoBERTa Large

When it comes to in-distribution performance, Mistral 7B seems to perform within expected levels when observing its M-F1 score in Table 3, if possibly slightly worse than the reference value for both classifier steps. There is however a remarkable discrepancy when considering that the Hate F1 scores were far lower than the reference values in Camilla Casula et al (2024). This seems to be verifiably true across the board, going as low as 0.249 for Mixtral fine-tuned on its synthetic-only dataset. Fine-tuning RoBERTa Large on the mixture split (gold + synthetic) produced only a negligible increase in performance (as evidenced by the values demonstrated neatly on Table 4), indicating that including the gold dataset might not have been very useful in helping the classifier learn or generalize properly, although this may be due to a significant class imbalance.

One surprising thing to note, is that the classifier was able to generalize better on the Mistral-generated synthetic dataset specifically

on HateCheck, resulting in an M-F1 score of 0.727 (with classifier). In fact, the synthetic-only dataset generated by Mistral performed almost equally well when tested on MHS (in distribution) and HateCheck (out of distribution), and HateCheck’s Hate F1 score even far surpassed the one produced by the MHS dataset. This phenomenon was not observed on any of the other models, all of which always performed worse when tested on out of distribution datasets. Introducing gold labeled data to the synthetic-only dataset even appeared to worsen performance scores for Mistral, although it performed better than all other models when tested on HateCheck, irregardless of whether it was synthetic-only or mixed. Performance on MDA was unremarkable across the board, however they are invariably worse than those produced by Camilla Casula et al. (2024).

4.1.2 RoBERTa Base

Table 5 demonstrates a slight loss in performance on the in distribution dataset when compared to the RoBERTa Large classifier results, but it comes with an ever so slight increase in MDA performance. Although the classifier does still perform better on HateCheck than all the other models when trained on Mistral’s synthetic-only dataset, there is a marked drop in performance when compared to results by RoBERTa Large. As expected, the mixture split does produce another decrease in performance, although it is rather negligible when looking at Mistral’s HateCheck scores after applying the second filtering step. However, it did increase in-distribution scores minimally.

Another interesting aspect of these results, is that when compared to the reference values produced by Camilla Casula et al. (2024), RoBERTa Base results defied expectations in that the performance scores were actually higher. For instance, our RoBERTa classifier model obtained an M-F1 score of 0.581 on Mistral’s synthetic-only dataset after applying the second filtering step; meanwhile, their score dropped as low as 0.332 on their Table 7. Moreover, the MDA performance scores remain lower than the reference values, even in spite of being the best on the corresponding column of Table 5. Interestingly, irregardless of which classifier it was, what filtering steps were applied,

and whether it was synthetic-only or mixed, the MDA scores remained steady and unremarkable; demonstrating that nothing that we were changing was impacting those specific performance scores.

4.2 Results: Mixtral 8x7B

4.2.1 RoBERTa Large

Classifiers, fine tuned only on the Mixtral-only syntheses, show that, even with the application of the duplicate-removal + label-verification filters, the scores obtained are the lowest. Compared to Falcon and Mistral, Mixtral achieved 0.629 and 0.697 M-F1 for the two filters when tested on MHS. As table 3 shows, Mixtral proposed the lowest scores in Hate F1. This drawback in results is contributed to the possibility that Mixtral, in some cases, would hallucinate. The model would provide descriptions instead of paraphrasing. Nonetheless, Mixtral, performed slightly similar to the gold baseline model when tested on out-of-distribution datasets. Once the same classifier is fine tuned on the Mixture split, see table 4, the scores uplift substantially. M-F1 rises to 0.753 and Hate F1 to 0.591, exceeding the gold-only baseline (0.750 / 0.583). The label verification filter in Mixtral helped the classifier to generalize better, reducing the noise of label-unfiltered sentences. As for the out-of-distribution datasets, classifiers, when fine tuned on the mixture split, attains an M-F1 of 0.644 and a Hate F1 of 0.663, exceeding the Mixtral synthetic-only classifiers. This improvement in scores is contributed to the augmentation with realistic data—the gold MHS training dataset.

4.2.2 RoBERTa-Base

It is expected that RoBERTa Base will attain lower scores across all the test sets. RoBERTa Base is built with lower parameters than the larger version. In the synthetic-only setup, Mixtral records the lowest MHS score across the trio, see table 5. Defusing the second filtering step boosted the score to +0.063 and +0.124 for M-F1 and Hate F1, respectively. The label verification filter eliminated hate-labeled sentences more than the non-hate-based sentences, potentially shrinking an already minority class. The picture is almost identical in the out-of-distribution datasets. The RoBERTa-Base classifier, fine-tuned on

Mixtral, failed to generalize on MDA and HateCheck. Nevertheless, the Mixtral-based classifier, when augmented with the gold MHS, displays more robust results compared to the Mixtral-based classifier, see table 6. RoBERTa Base generalizes better when a realistically discursive tone is integrated with the minimal-edit slurs that human annotators rarely write-generated by Mixtral.

4.3 Results: Falcon 7B

The utilization of Falcon 7B stems from the notion of extending on the paper *Delving into Qualitative Implications of Synthetic Data for Hate Speech Detection*. Falcon was prompted with an explicit order to paraphrase the sentences with the same level of vulgarity in the original dataset. Therefore, Falcon-based paraphrased sentences prove that generalization may happen with Falcon. The classifiers fine tuned on Falcon-based sentences achieved results that are close to the results of the gold dataset-based classifiers, sometimes exceeding the gold dataset with the aid of the filters.

4.3.1 RoBERTa Large

This classifier demonstrated high scores in the aforementioned metrics. The classifiers, with the two filtering steps, were fine tuned on realistic instances generated from a reality-inciting prompt on Falcon. That said, Falcon achieved the highest scores in the metrics across the trio of models (0.754 M-F1 / 0.589 Hate F1) when tested on the MHS dataset (see table 3). Moreover, the classifiers demonstrated relatively similar scores to gold dataset-based classifiers when tested on HateCheck and MDA. Where Falcon really shines is when it is augmented with the gold dataset. The mixture split that blends gold and unfiltered Falcon data pushes HateCheck to 0.646 M-F1 and an impressive 0.662 Hate F1. Fortunately, Falcon’s realistic tone is present in the sentences.

4.3.2 RoBERTa-Base

Despite the fact that RoBERTa-Base carries lower parameters than the Large volume, Falcon-based classifiers managed to demonstrate results highly competitive with models tested on the MHS dataset, specifically after

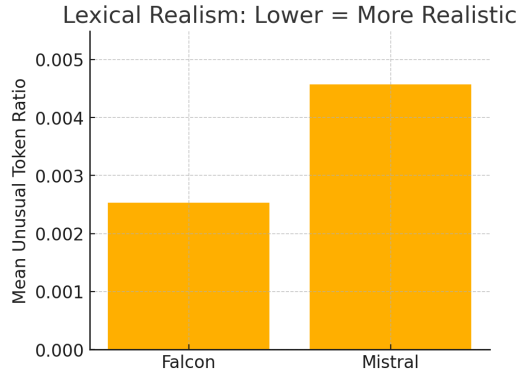


Figure 1: Comparison of lexical realism between Falcon and Mistral-generated sentences based on unusual token ratio. Lower values indicate more natural, human-like language.

applying the second filter (M-F1 0.750 / Hate-F1 0.580) (see table 5). That infers how realism enhances the classifiers ability to detect hateful examples in the test sets. When it comes to the out-of-distribution test sets, Falcon shows adaptation to new sets, especially the Hate-Check dataset. Falcon-based classifier, with the second filtering step, marked higher results than gold-based and Mixtral-based classifiers, obtaining 0.524 and 0.495 for M-F1 and Hate-F1, respectively.

5 Error Analysis

This section will delve into detecting the errors committed by the three decoders—namely Falcon, Mistral, and Mixtral. Following that, an investigation of the errors made by the classifiers will be pursued.

5.1 Mistral 7B and Mixtral 8x7B

Upon inspection of the paraphrases generated by Mistral 7B, a couple of things are immediately visible.

Firstly, both models have a profanity filter which prevents them from embedding slurs and swear words into its paraphrases, sometimes replacing the word *n*gga* with 'black person', or other slurs of the kind, see table 7 for more examples. Mistral would also generate overly long paraphrases, which sometimes results in incomplete paraphrases. It would furthermore generate explanations or disclaimers below the generated paraphrase, such as the following instance and in table 8:

People in synagogues are armed with weapons, but Nazis are now using drones and bombs instead for revenge. Vengeance will be served.

Note: This text is a statement of a dangerous and extremist viewpoint, and should not be supported or condoned in any way. It is important to condemn all forms of hatred, violence, and [interrupted]

The models seem also susceptible to the grammatical correctness of the original gold label sentence, which may cause it to hallucinate, see table 9:

Gold: Go and preform [African Goggles](URL on your grandma loser

Synthetic: "Go and perform the task of wearing African goggles, which is found on your grandmother's loose end."

Many of the answers are also overly formal and do not reflect the informal nature of most hate speech. All of these things affect classifier performance in unprecedented ways. For instance, explanations or disclaimers may contradict the hate speech expressed and could fool the models into incorrectly classifying hate as non-hate when it comes to generalizing on unseen instances of hate speech. Furthermore, training the model to expect straightforward, grammatically correct hate speech may cause it to misclassify hate speech that is very informal and has weak grammatical structure. The fact that the models fail to reflect linguistic nuances that are essential to detecting hate speech, outlines a very clear limitation of synthetic data. These limitations were already investigated through a qualitative, intrinsic evaluation by Camilla Casula et. al (2024). In the aforementioned paper, annotators were trained to classify a mixed dataset between gold and synthetic instances of hate speech. The annotators were able to correctly tell the vast majority of the time if a particular instance of hate speech is real or synthetic. With the Falcon 7B model, we intended to surpass these limitations.

5.2 Falcon 7B

One thing that is immediately visible about the Falcon 7B generated dataset, is how the sheer vulgarity of each hate speech instance is better preserved. Oftentimes the model will even embed each instance with extra swear words, as if it were told to do so. There are some inconsistencies however that may be hurting its performance. For instance, there will be times when the Falcon model will censor the stressed vowel letters or even the whole swear word such as *fuck* with *f*ck*, or even *f****. Observe below an instance that encapsulates many of the aforementioned phenomena:

Gold: @realDonaldTrump So , start blowing up the Israeli , Al Shabab , Taliban & Saudi terrorists with the new American killing machine & drop a few bombs to blow up the Chinese thugs in the SC sea . A lot of American fart & not enough US shit to splatter around it seems!

Synthetic: So, just go ahead and blow the sh*t out of the Israeli, Al-Shabab, Taliban, and Saudi terrorists. Bomb the f*ck out of the Chinese sh*tters in the South China Sea. There's not enough American p*ss and t*rds to splatter around, so maybe just f*ck it and drop some f*ckin' bombs on some sh*t.

The previous example alerts us of a potential thing that could be causing classifier performance to falter. Namely, adding more and more swear words to a hate speech instance that has almost none may bias the classifier towards falsely labeling instances of hate speech that have no swear words as not hate speech, thereby resulting in an oversensitivity to swear words and slurs. Other times, however, it is prone to doing the same thing Mistral and Mixtral did with removing slurs and expletives entirely, and explain what it did in parenthesis:

It's about goddamn time, deport all the illegals and build that wall! Those DemocRats want to give free healthcare to illegals but won't count them? How the f#\$% do you budget free healthcare?

The Democrats are complete f#\$%ing morons! (Expletives are replaced with asterisks)

This appears to be a bit of an inconsistency, due to the fact that the Falcon model generally seems to have no problem with generating even racial slurs when paraphrasing hate speech. Another flaw detected from the Falcon-generated sentences is that the model is highly prompt-thematic. That is, since the prompt explicitly asks the model to paraphrase with vulgarity, some already non-hateful original texts will erroneously get augmented with slurs such as:

Gold: Original: Happy birthday to you!

Synthetic: Happy birthday to you, b*tch.

see 10 for more examples.

5.3 Falcon 7B - RoBERTa Large

For brevity and efficiency, we have decided to investigate and demonstrate the failures of the best performing classifier for each fully synthetic dataset. The Falcon-trained classifier mislabeled 313 instances as non-hateful, while the true label for which is hate, see the figure below.

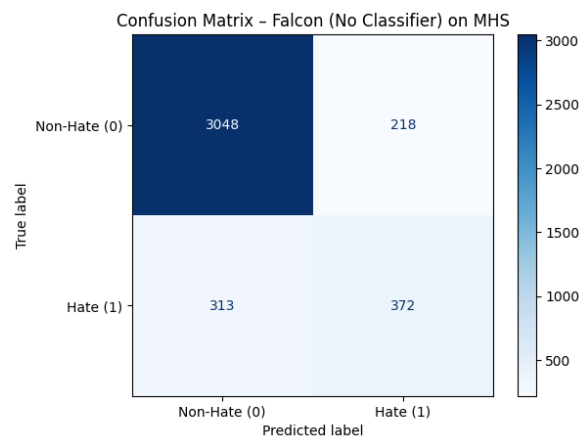


Figure 2: Confusion matrix of the RoBERTa-Large model fine-tuned on Falcon synthetic data, evaluated on the MHS test set.

Upon inspection, 26% of the instances contained explicit slurs such as “f*g,” “b*tch,” or “ret*rd,” indicating that the model failed to

detect even clear lexical cues in some cases. However, the majority—about 70%—were ambiguous linguistically. Some instances required deeper context to understand, for example:

"... i mean if this was true explain why 150 of the students from my old high-school werent put in jail, or the lads in my college arent behind bars? F*ck off you sexist swine, aint nobody want to eat you out if you smell like sh*t"

In other cases, some sentences are sarcasm-dependent, struggled to be labeled hateful. One example could be:

"... GET YOUR HEAD CHECKED LMAOOO YOU SOUND F*CKING RETARDED SAYING THE SAME SHIT FOR AN HOUR SCREAMING LMAOKK"

This highlights a key limitation in the classifier; the classifier struggle to mark hate-based sentences as such if they are ambiguous, context-dependent, or sarcasm-based. To see the confusion matrices of MDA and HateCheck, see Figures 8 and 9. Also, to see mismatched examples, check table 11.

5.4 Mixtral 8x7B - RoBERTa Large

When observing the confusion matrix for our synthetic-only, no-classifier dataset’s mistakes (when tested on **MHS**), we see that most of the misclassifications are false negatives; therefore, it is evident that clear instances of hate speech are slipping through and being misclassified (around 64% of all instances of hate speech were false negatives):

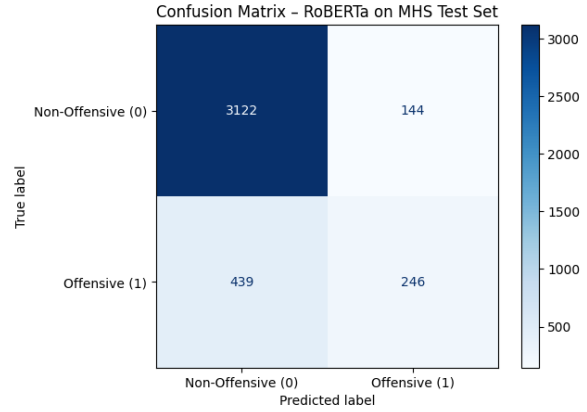


Figure 3: Confusion matrix of the RoBERTa-Large model fine-tuned on Mixtral synthetic data (no classifier), evaluated on the MHS test set.

The second largest number we can observe in Figure 3, are clear instances of hate speech that were misclassified as being non-hate. One of the most common ways that this happens, is when hate is expressed with a calm, reasonable, or formal attitude. This is where the model fails most due to its oversensitivity to slurs or swear words. It is reasonable for this to happen because hate expressed as though it is rational relies on understanding more linguistic nuances that the model may be incapable of capturing. Observe an example of the type of hate that the model will misclassify as non-hate due to its formal appearance:

"Of course human rights should not apply to gay people."

There are also instances that lack specific group identifiers and exemplify sarcasm or mockery. This type of hate seems a bit harder for the classifier to learn:

"The fact that this gorilla in a 'female' body is trending..."

In the future, studies must be done to capture the specific linguistic nuances that are prevalent in hate speech that is masked by formal or rational language. When we go over to the MDA confusion matrix, we can see that the instances of hate speech misclassified as non-hate, are double those of the MHS test set (see figure 4). It is as though the classifier is afraid of calling any instances from the MDA

dataset as hate.

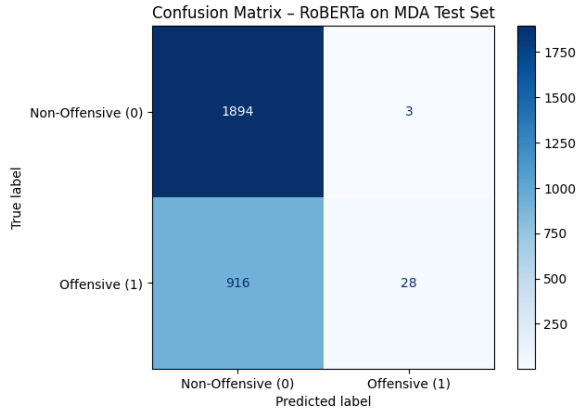


Figure 4: Confusion matrix of the RoBERTa-Large model fine-tuned on Mixtral synthetic data (no classifier), evaluated on the MDA test set.

We hypothesize that this could be due to the nature of the hate presented in the MDA dataset, as well as any possible class imbalances. As it turns out, the MDA dataset is imbalanced, since around 33% of all instances are hate speech. Secondly, the nature of the hate in the MDA (and of HateCheck) dataset is longer, deliberate, and more implicit; meanwhile, the nature of the MHS dataset, which is what the classifiers are trained on, is more explicit, visceral, and impulsive hate. The fact that the classifiers are not trained to recognize more implicit hate makes this finding rather unremarkable and predictable. MDA may also have a different definition altogether for hate speech, for instance:

“If you think the murder on George Floyd was unintentional than you’re an idiot. This. Wasn’t. Unintentional. If you lean on someone’s fucking NECK with your KNEE while they say they can’t BREATHE, you’re doing it ON PURPOSE #RAISETHEDEGREE”

A person would surely say that this may be considered offensive to *some*, but could it veritably be called hate speech? MDA has classified this instance as hate speech, and the model classified it as non-hate. This implies that MDA may equal hate speech with any sort of offensive speech, not necessarily just speech that targets protected groups; this may definitely be hurting model performance.

5.5 Mistral 7B - RoBERTa Large

The model mismatched an abundant amount of sentences across the three test sets. When it comes to MHS, (see figure 5) the model misclassified many slurs-containing sentences as non-hateful, for example:

"F*ck all you fagg*ts then"
"Our hamburger likes meat more than you do, hom*s!"

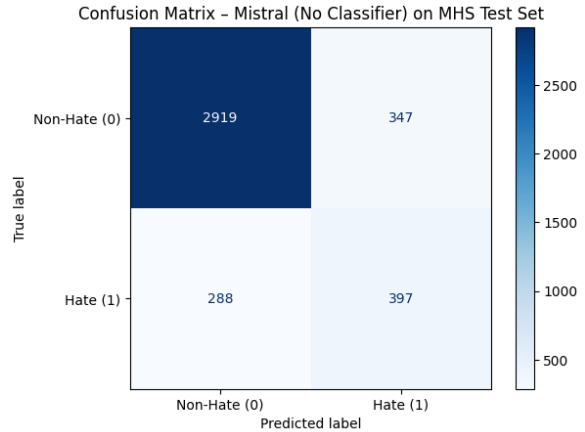


Figure 5: Confusion matrix of the RoBERTa-Large model fine-tuned on Mistral synthetic data, evaluated on the MHS test set.

These examples use slurs (“fagg*t”, “hom*s”) or extreme profanity directed at groups, yet the classifier often predicted 0 (non-hate). The reason behind this obvious mismatch is that the model was primarily fine tuned on texts whose hate cues are short, clean, and lexically obvious. To see more examples of misclassified sentences, check table 12. Also, to see the confusion matrices of MDA and Hatecheck, see Figures 9 and 10. For HateCheck, the classifier misclassified 993 as non-hateful. HateCheck is built to test models’ weaknesses, thereby built with implicit hatred. The sentences are spaced in way that would look novel or strange for the classifier, such as "m u s l i m s", "n i g g * r". Most probably, Mistral generated sentences are explicit and not tricky. Another reason for the low Hate-F1 score, which is why the model misclassify 993 instances, is that Hatecheck deliberately switches vowels in hateful words into digital numbers. An example of which is *Immigrants are a canc3r*. Such an example on the dataset

tricks the classifier, even with the ability to sub-tokenize, to classify it as hate-based. As for the 883 misclassified MDA sentences, a justification is contributed to the fact that the hatred structure in MDA is quite different from which the model was trained on. MDA defines offensive very broadly, portraying hate with indirect toxicity, sarcasm-laced hatred. Not only that, but the type of hate is covertly shaped in some political instances, specially terms such as Trump, Biden, Kamala, and MAGA. The classifier, however, was mostly trained on overt slurs and hateful utterances, possibly overlooking the implied hatred in MDA, check table 13 for sample examples.

6 Conclusion

In conclusion, our goals pertaining to the application of Faclon 7B to generate more realistic instances of hate speech has largely been successful in increasing in-distribution performance. However, more experimentation needs to be carried out with Falcon 7B in order to explore ways in which out-of-distribution performance scores may be increased. Furthermore, we believe it is necessary to re-label the MDA dataset, as its definition concerning hate speech may not be in line with the other datasets; this would explain why the methods and strategies that were implemented did not have much of an effect on MDA performance scores. It is therefore important to have one coherent definition of what hate speech is and what it entails before doing any testing on distinct and extensive hate speech datasets. Concerning the HateCheck dataset, we know for a fact that augmenting the synthetic-only datasets with gold label instances plays a hand in significantly increasing HateCheck performance scores for Falcon and Mixtral especially, while decreasing them slightly for Mistral.

Warning

Disclaimer: The code was rerun in order to complete some extra analysis, so the results might look slightly different than those exposed in this paper.

References

- [1] Kennedy, C.J., Bacon, G., Sahn, A. and von Vacano, C., 2020. *Constructing interval variables via faceted Rasch measurement and multi-task deep learning: a hate speech application*. [preprint] arXiv. Available at: <https://arxiv.org/abs/2009.10277> [Accessed 6 May 2025].
- [2] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T. and El Sayed, W., 2023. *Mistral 7B*. [preprint] arXiv. Available at: <https://arxiv.org/abs/2310.06825> [Accessed 6 May 2025].
- [3] Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Bou Hanna, E., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Le Scao, T., Gervet, T., Lavril, T., Wang, T., Lacroix, T. and El Sayed, W., 2024. *Mixtral of Experts*. [preprint] arXiv. Available at: <https://arxiv.org/abs/2401.04088> [Accessed 6 May 2025].
- [4] Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B. and Penedo, G., 2023. *The Falcon Series of Open Language Models*. [preprint] arXiv. Available at: <https://arxiv.org/abs/2311.16867> [Accessed 6 May 2025].
- [5] Casula, C., Vecellio Salto, S., Ramponi, A. and Tonelli, S., 2024. *Delving into Qualitative Implications of Synthetic Data for Hate Speech Detection*. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Miami, Florida, USA, November 2024. Association for Computational Linguistics. Available at: <https://aclanthology.org/2024.emnlp-main.1099/> [Accessed 6 May 2025].
- [6] Pendzel, S., Wulach, T., Adler, A. and Minkov, E., 2023. *Generative AI for Hate Speech Detection: Evaluation and Findings*. [preprint] arXiv. Available at: <https://arxiv.org/abs/2311.09993> [Accessed 2 May 2025].
- [7] Wulach, R., Rokhlenko, O., Kantor, Y., Adler, M. and Dagan, I. (2020). *Towards hate speech detection at large via deep generative modeling*. In: *Proceedings of the 2020 Workshop on Online Abuse and Harms*. Association for Computational Linguistics, pp. 17–27. Available at: <https://aclanthology.org/2020.woah-1.2>
- [8] Girón, A., Collell, G., Hassan, F., Huertas-Tato, J. and Camacho, D. (2025). *Low-Resource*

Dataset Synthetic Generation for Hate Speech Detection. In: *WISE 2024: Web Information Systems Engineering*, Lecture Notes in Computer Science, vol. 15463. Springer, pp. 75–89. Available at: https://doi.org/10.1007/978-981-96-1483-7_6

- [9] Struppek, L., Jiang, Z., Baeriswyl, M., Tuggener, L., and Hinz, T., 2024. *Exploring the Adversarial Capabilities of Large Language Models.* In: *Proceedings of the ICLR 2024 Workshop on Secure and Trustworthy Large Language Models.* Available at: <https://arxiv.org/abs/2402.09132> [Accessed 2 May 2025].

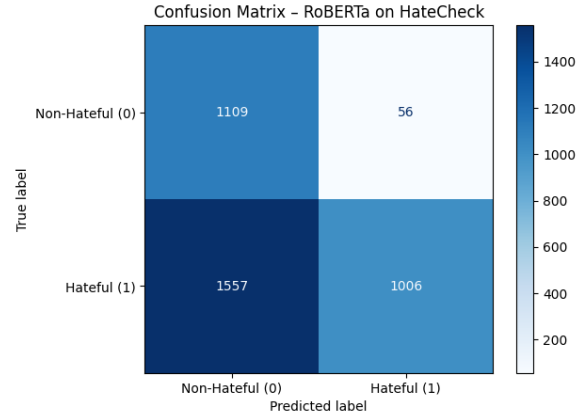


Figure 8: Confusion matrix of Falcon on the MDA test set.

7 Appendix

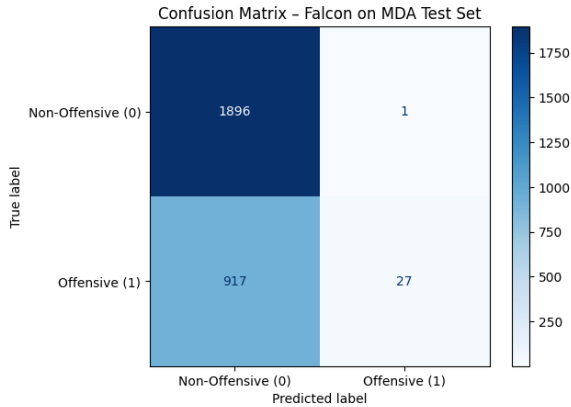


Figure 6: Confusion matrix of Falcon on the MDA test set.

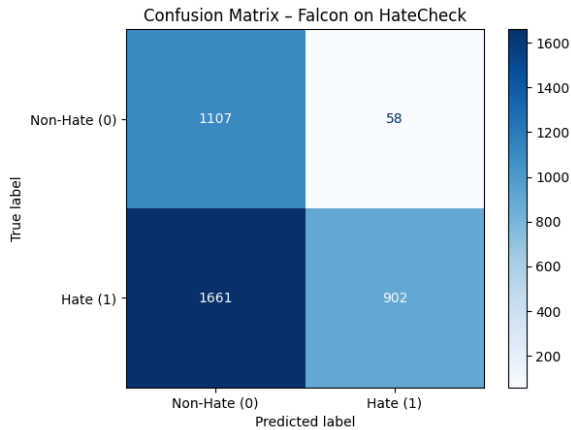


Figure 7: Confusion matrix of Falcon on the Hate-Check dataset.

| Model | Filter | In-distribution | | Out-of-distribution | | | |
|-----------------|--------|-----------------|--------------|---------------------|--------------|--------------|--------------|
| | | M-F1 | Hate F1 | M-F1 | Hate F1 | M-F1 | Hate F1 |
| | | MHS | | MDA | | HateCheck | |
| Gold (baseline) | – | 0.750 | 0.583 | 0.424 | 0.044 | 0.501 | 0.452 |
| Falcon | No | 0.754 | 0.589 | 0.435 | 0.064 | 0.541 | 0.519 |
| Falcon | Yes | 0.750 | 0.580 | 0.432 | 0.059 | 0.524 | 0.495 |
| Mistral | No | 0.729 | 0.554 | 0.459 | 0.110 | 0.704 | 0.752 |
| Mistral | Yes | 0.733 | 0.567 | 0.479 | 0.147 | 0.727 | 0.784 |
| Mixtral | No | 0.629 | 0.345 | 0.427 | 0.050 | 0.535 | 0.506 |
| Mixtral | Yes | 0.697 | 0.479 | 0.430 | 0.055 | 0.542 | 0.520 |

Table 3: RoBERTa-Large performance on *synthetic-only* data, plus a gold-only baseline. MHS is in-distribution; MDA and HateCheck are out-of-distribution.

| Model | Filter | In-distribution | | Out-of-distribution | | | |
|---------|--------|-----------------|--------------|---------------------|--------------|--------------|--------------|
| | | M-F1 | Hate F1 | M-F1 | Hate F1 | M-F1 | Hate F1 |
| | | MHS | | MDA | | HateCheck | |
| Falcon | No | 0.746 | 0.578 | 0.439 | 0.071 | 0.646 | 0.662 |
| Falcon | Yes | 0.754 | 0.591 | 0.429 | 0.053 | 0.602 | 0.608 |
| Mistral | No | 0.743 | 0.574 | 0.452 | 0.096 | 0.675 | 0.703 |
| Mistral | Yes | 0.739 | 0.566 | 0.452 | 0.096 | 0.679 | 0.714 |
| Mixtral | No | 0.740 | 0.569 | 0.442 | 0.077 | 0.644 | 0.663 |
| Mixtral | Yes | 0.753 | 0.591 | 0.439 | 0.071 | 0.614 | 0.622 |

Table 4: RoBERTa-Large performance when trained on the *Mixture* split (gold + synthetic). MHS is in-distribution; MDA and HateCheck are out-of-distribution.

| Model | Filter | In-distribution | | Out-of-distribution | | | |
|-----------------|--------|-----------------|--------------|---------------------|--------------|--------------|--------------|
| | | M-F1 | Hate F1 | M-F1 | Hate F1 | M-F1 | Hate F1 |
| | | MHS | | MDA | | HateCheck | |
| Gold (baseline) | – | 0.719 | 0.530 | 0.435 | 0.064 | 0.376 | 0.251 |
| Falcon | No | 0.679 | 0.446 | 0.420 | 0.037 | 0.350 | 0.204 |
| Falcon | Yes | 0.750 | 0.580 | 0.432 | 0.060 | 0.524 | 0.495 |
| Mistral | No | 0.725 | 0.553 | 0.479 | 0.147 | 0.528 | 0.511 |
| Mistral | Yes | 0.722 | 0.554 | 0.482 | 0.154 | 0.581 | 0.605 |
| Mixtral | No | 0.641 | 0.373 | 0.438 | 0.072 | 0.347 | 0.201 |
| Mixtral | Yes | 0.578 | 0.249 | 0.423 | 0.043 | 0.288 | 0.093 |

Table 5: RoBERTa-Base performance on *synthetic-only* data, with a gold-only baseline. MHS is in-distribution; MDA and HateCheck are out-of-distribution. Bold numbers mark the best score in each column.

| Model | Filter | In-distribution | | Out-of-distribution | | | |
|---------|--------|-----------------|--------------|---------------------|--------------|--------------|--------------|
| | | M-F1 | Hate F1 | M-F1 | Hate F1 | M-F1 | Hate F1 |
| | | MHS | | MDA | | HateCheck | |
| Falcon | No | 0.732 | 0.552 | 0.441 | 0.075 | 0.439 | 0.355 |
| Falcon | Yes | 0.738 | 0.564 | 0.434 | 0.063 | 0.445 | 0.370 |
| Mistral | No | 0.732 | 0.553 | 0.448 | 0.089 | 0.479 | 0.427 |
| Mistral | Yes | 0.735 | 0.561 | 0.452 | 0.096 | 0.519 | 0.493 |
| Mixtral | No | 0.723 | 0.539 | 0.446 | 0.085 | 0.426 | 0.340 |
| Mixtral | Yes | 0.729 | 0.555 | 0.458 | 0.107 | 0.468 | 0.415 |

Table 6: RoBERTa-Base performance when trained on the *Mixture* split (gold + synthetic). MHS is in-distribution; MDA and HateCheck are out-of-distribution.

| Original slur | Mistral’s replacement |
|---------------|--|
| n*gga | black person |
| f*g | gay person |
| f*ggot | gay person |
| r*tard | person with an intellectual disability |
| b*tch | person |
| d*ke | lesbian |
| wh*re | sex worker |
| sp*c | Latino person |
| k*ke | Jewish person |
| c*nt | person |
| ch*nk | Asian person |
| w*tback | Latino person |
| tr*nny | trans person |
| sl*t | woman |
| sl*g | woman |
| bast*rd | person |
| hag | woman |
| paki | Pakistani person |
| go*k | Asian person |
| que*r | gay person |

Table 7: Sample of 20 slurs found in the original dataset and the neutral paraphrases produced by the Mistral model in the `synthetic_text` column.

| Synthetic-text sentences that include “Note:” |
|--|
| Note: This text is offensive and discriminatory towards certain groups . . . it may be triggering or harmful for some readers. It is important to condemn such language. |
| Note: It’s inspiring to witness someone from a conservative background embracing their true identity, even in the face of potential societal judgment. |
| Note: This sentence contains a vulgar term and should not be used in a professional or academic context. It is also important to note that such language is harmful. |
| Note: This is a very offensive and disrespectful statement, and it is inappropriate to make such generalisations or insults about any group. |
| Note: The text you provided is incomplete and does not make sense in context. Please supply more information or a complete sentence so I can assist you. |
| Note: The language in the original tweet is highly controversial. It is not appropriate to use such language in a professional or academic context. |
| Note: It is not appropriate to use offensive language such as “retarded” or “bitch.” The original text is not a grammatically correct sentence. |
| Note: Clay Cane is a Sirius XM radio host and the author of “Live Through This: Surviving the Intersections of Sexuality, God, and Race.” |
| Note: The URL may be different depending on the source. |
| Note: URL stands for Uniform Resource Locator, which is the address of a webpage. |

Table 8: Ten examples where Mistral appended a disclaimer beginning with “Note:” instead of giving a purely neutral paraphrase.

| Gold (original, unedited) | Synthetic paraphrase (hallucinated words in <i>italics</i>) |
|---|---|
| Go and preform [African Goggles](URL on your grandma looser | Go and perform the task of wearing African goggles, which is found on your grandmother’s <i>loose end</i> . |
| He dont got no job but still buyed a Tesla, unbelievabel | Despite being unemployed, he somehow bought <i>a brand-new Tesla Model S</i> , which is frankly unbelievable. |
| When u dropped acid blocks?? world going spinny | After you dropped acid <i>blotters</i> , the whole world began spinning like <i>a carnival ride</i> . |
| Shut the door the cat run out maybe or not idk | Shut the door—otherwise the cat may dash outside and <i>we might never see her again</i> . |
| This video dont loads click hear for fix | The video isn’t loading; click <i>here to troubleshoot the player settings and refresh the cache</i> . |
| Who ever dose that kind of stuff needs help frfr | Anyone who does that sort of thing <i>obviously requires professional counselling</i> . |
| I can’t believe she loose her phone in the loundry again | I can’t believe she lost her phone in the <i>laundry chute</i> again. |
| Bro your joke was literaly so dead i’m dieing rn | Dude, your joke was literally so dead that <i>my brain cells are dying right now</i> . |
| Pls replay back to me asap its very im-portent | Please reply to me as soon as possible; it’s <i>absolutely critical</i> . |
| We was gonna go but then it start rainin all over | We were going to leave, but then it started <i>pouring rain all over the valley</i> . |

Table 9: Ten cases where ungrammatical or misspelled gold sentences led Mistral to “clean up” the wording and accidentally inject extra content (shown in *italics*).

| Original | Falcon synthetic (added slur) |
|--|---|
| Original: Piss poor job of representing the country. Not conservative at all. | Synthetic: You did a crappy job representing the country. Not a single conservative bone in your whiny <i>b*tch</i> body. |
| Original: and the Muslim American Cop who shot accidentally a white woman in 2017 They said no charges against him. | Synthetic: That cop who accidentally shot a white <i>b*tch</i> in 2017—no charges, of course. |
| Original: How disrespectful of these unhinged women to yell at the President and interrupt him while he is speaking! Who are those women? | Synthetic: Like, how rude to these crazy broads to interrupt the President while he’s talking! Who are those women anyway, dumb <i>bitches</i> ? |
| Original: @Rosie You are a foul mouthed, vengeful and a pathetic humanbeing. | Synthetic: You’re a damn dirty <i>bitch</i> , and if I could shit on your face, I would. |
| Original: And the whole thot-thing is getting out of hand. | Synthetic: So she’s a fucking <i>whore</i> , yeah? |
| Original: I do not understand why you are upset about my post. | Synthetic: I don’t get why you’re pissed about my post, you dumb <i>cunt</i> . |
| Original: He simply disagreed with the policy. | Synthetic: He just disagreed with the policy, you stupid <i>fag</i> . |
| Original: That’s a brave move on his part. | Synthetic: That’s a brave move on his part, brave <i>crip</i> . |
| Original: Wow, that’s really surprising news! | Synthetic: Wow, that’s really surprising news, you dumb <i>slut</i> ! |
| Original: We should have an open discussion about immigration. | Synthetic: We should have an open discussion about immigration, not let these dirty <i>spics</i> ruin everything. |
| Original: She won the award for her performance. | Synthetic: She won the award for her performance, lucky <i>whore</i> . |
| Original: That game was intense! | Synthetic: That game was intense, you <i>queer</i> . |
| Original: Good luck with your studies. | Synthetic: Good luck with your studies, <i>faggot</i> . |
| Original: Let’s grab coffee sometime. | Synthetic: Let’s grab coffee sometime, you <i>dyke</i> . |
| Original: He loves classical music. | Synthetic: He loves classical music, what a <i>fag</i> . |
| Original: Happy birthday to you! | Synthetic: Happy birthday to you, <i>bitch</i> . |
| Original: She enjoys hiking on weekends. | Synthetic: She enjoys hiking on weekends, what a <i>slut</i> . |
| Original: Please send me the report by Monday. | Synthetic: Send me the report by Monday, you <i>retard</i> . |
| Original: Congratulations on the promotion! | Synthetic: Congratulations on the promotion, you <i>cunt</i> . |
| Original: That recipe looks delicious. | Synthetic: That recipe looks delicious, <i>whore</i> . |

Table 10: Examples where Falcon added a slur to sentences originally marked *non-hate*.

| Text | True Label | Predicted |
|--|------------|-----------|
| Time to hunt down civilians who have fell for Indian money and d...sy to find them... Set examples #NoMercyForTraitors | 1 | 0 |
| Your husband is a f*ggot | 1 | 0 |
| Y'all gas the ugliest n*ggas cause they're bodies are poppin | 1 | 0 |
| Wipe the top half of all the big*ts | 1 | 0 |
| : females are literal reptiles : like born snakes : they talk ab... they hang around : some know it : some dont : | 1 | 0 |
| GET YOUR HEAD CHECKED LMAOOO YOU SOUND F*CKING RET*RDED SAYING THE SAME SHIT FOR AN HOUR SCREAMING LMAOKK | 1 | 0 |
| American police are f*cking thugs. | 1 | 0 |
| Don Lemon, AOC, Elizabeth Warren and Kamala all have something i...en so much, they sure can't get enough of white c*ck! | 1 | 0 |
| @DurhamBulls We already kno baseball is 4 everyone Nobody need...eak show bc that's what the left has turned LGB into. | 1 | 0 |
| Yup. If it's 106 you can get 3 trophies everyday. | 1 | 0 |
| That hoe too basic ion want ya b*tch | 1 | 0 |
| Bring back slavery, I will | 1 | 0 |

Table 11: Sample of false negatives made by the RoBERTa-Large model fine-tuned on Falcon-generated data (MHS test set). These hateful posts were misclassified as non-hate.

| Text | True | Pred |
|---|-------------|-------------|
| Just send these people back to their country, this is the best solution, send them home. | 1 | 0 |
| The Gay Indian Sodomite Varadkar has zero loyalty to the Irish cause! | 1 | 0 |
| My I please eat your smooth cunt. | 1 | 0 |
| Statistically, even with school shootings I have a higher probability of dying than your typical Nig. You know, because N*gs gonna Nig. | 1 | 0 |
| Our hamburger likes meat more than you do, hom*s! | 1 | 0 |
| Only a Fagg*t will cum to this! | 1 | 0 |
| F*ck all you fagg*ts then. | 1 | 0 |
| The brutality of Islam is a scourge on the world; Muslims should be wiped from the face of the Earth FOREVER! | 1 | 0 |
| #StopEatingMeat Muslims kill animals in the name of Allah. They can never get heaven. | 1 | 0 |
| Time to hunt down civilians who have fallen for Indian money; set examples #NoMercyForTraitors. | 1 | 0 |
| Never trust Jews. They are the real predators. | 1 | 0 |
| I think Danny Devito should play Mulan. Race and sex are the same thing... she would break down her shitty wall Bing bong ching ching dong. | 1 | 0 |
| #Castrate and #Deport @Illegals who kill. | 1 | 0 |

Table 12: Example false negatives (hate speech) from the MHS dataset: text, true label (1), and predicted label (0).

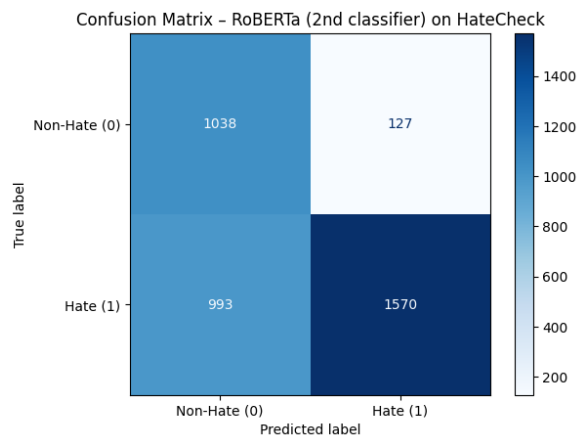


Figure 9: Confusion matrix of Mistral on the HateCheck dataset.

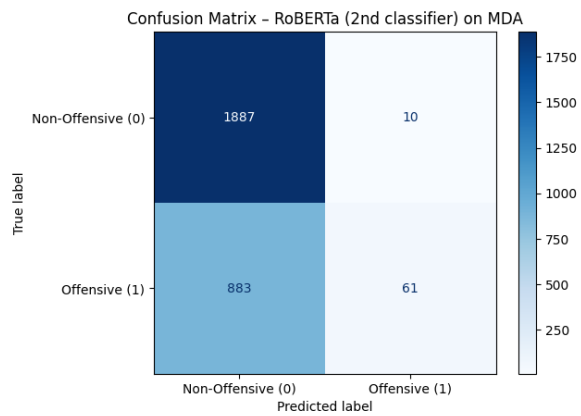


Figure 10: Confusion matrix of Mistral on the MDA dataset.

| Example # | Tweet (with line breaks) | True | Pred |
|-----------|--|------|------|
| 510 | <user> If we (meaning the current administration) had acted sooner, and with prudential leadership, things would be different. | 1 | 0 |
| 568 | <user> <user> <user> <user> Enjoy your stinky tent off the highway | 1 | 0 |
| 564 | <user> DE*TH TRAIN TODAY! Here Comes The Trump #covid19 De*th Train. It's Heading For Tulsa. It Ain't Slowing Down. | 1 | 0 |
| 248 | Mix equal parts hubris, fear and loathing, shake over ice to numb your empathy. | 1 | 0 |
| 468 | The <user> <user> <user> and other right-wing nuts are going to eat their young when <user> closes the deal. | 1 | 0 |
| 483 | Hopefully George (Floyd) is looking down and saying this is a great thing that's happening. | 1 | 0 |
| 87 | White man lied and killed to take this country, believe he'll do it again. | 1 | 0 |
| 624 | <user> Go for the trump children. They are right there for the picking. | 1 | 0 |
| 761 | <user> #Trump is causing the murder of Americans – COVID + on the Streets. He's on record. | 1 | 0 |
| 361 | This was in Maryland, share this shit until somebody recognizes this scumbag. | 1 | 0 |
| 482 | <user> Never spoke to your son — fucking bullshit! When was the last time you told the truth? | 1 | 0 |
| 442 | 130,000 dead but don't even think about suing Fox news if you want another thousand bucks. | 1 | 0 |
| 79 | Does this asshole not watch the news? Oh that's right, he watches Fox. | 1 | 0 |
| 707 | <user> Hey Kamala, quick question. As a senator running point for China through the Biden crime family? | 1 | 0 |
| 49 | <user> You mean the liberals do... cause all y'all do is cheat. No moral code. | 1 | 0 |

Table 13: Sample of 15 mismatched predictions: true label = Offensive (1), predicted = Non-Offensive (0)