AltBERT: Domain Specific Pretraining on Alternative Social Media to Improve Hate Speech Classification

Stanford CS224N Custom Project

Jackson Eilers, Oliver Brady, and Julia Wang
Stanford University
jeilers@stanford.edu, oqbrady@stanford.edu, jwang00@stanford.edu

Abstract

Fringe online communities like 4chan and Parler are havens of hate speech and tend to develop unique vocabularies that are challenging for hate speech classifiers to decode. These vitriolic environments normalize hateful dialogue that has proven to elicit real world violence, making hate speech classification an important issue for both online and offline safety. We aim to perform hate speech classification on three domain specific hate speech datasets from Twitter [1], Reddit [2], and Gab [2]. Our aim is to improve results of hate speech classification within these fringe communities by using transfer learning by pretraining BERT models on domain specific corpora from Parler and 4chan. We build off related works by using the BERT base uncased model as our baseline. We contribute to these works by pretraining the BERT model on larger corpora and corpora of supposedly similar domains to the finetuning datasets to test whether larger and similar pretraining datasets improve results. We also modified the domain specifc exBERT model [3] for hate speech classification. The models pretrained on a large corpus of Parler and 4chan posts showed accuracy improvements over the baseline in two of the three hate speech datasets. Improvements were observed in datasets of a similar domain to our pretraining corpora. Our exBERT model did not show improvements to the baseline due to limitations in the existing exBERT codebase.

1 Key Information to include

Our mentor is Anna Yang. We have no external collaborators and are not project sharing.

2 Introduction

Alternative social media sites like 4chan, Parler, and Gab advertise as online havens of free speech. Consequently, the lack of content moderation in the name of free speech drives a slew of racist, sexist, and other hateful content to flood their sites. This toxic environment can be harmful to other individuals online and normalizes hateful dialogue. In many countries, hate speech is a felony, making the detection of it important to accurately classify. Notably, hate speech is protected under the First Amendment in the US. Many papers have focused on developing hate speech classification models ([1], [2], [4]) but few have focused on hate speech detection in insular online communities like 4chan and 8kun. These communities often develop unique, coded vocabularies which poses a problem for hate speech classifiers. The January 6th insurrection, an event Parler users referred to as "the Storm," illustrates their use of coded language. Despite observing significant improvement in recent years due to the powerful classification abilities of transformer architectures like BERT, hate speech classifiers still struggle to classify this kind of hateful coded language. Building classifiers that are able to pick up on this domain specific vernacular will be of increased importance given the rise of real world violence stemming from these sites including acts of terrorism like the Christchurch massacre [5] which originated on 8chan and the insurrection against the US Capitol on Jan. 6, 2021 [6] which originated on Parler. The difficulties of hate speech classification include classifying the type of hate speech (sexism, racism, homophobia, etc.), the severity of the speech (hateful, offensive, etc.) [1], detecting hate speech in various languages [7], and detecting coded, domain specific hate speech, to name a few. Recent hate speech classification models show BERT models obtain state-of-the-art results for hate speech classification [8], which served as the impetus for using the BERT base uncased model as a baseline in our experiment. Other methods have shown both the successes and limitations of pretraining BERT on domain specific corpora for hate speech classification [9]. Transfer learning

from pretraining performs equal to or worse than the baseline BERT models due to limitations in the datasets. We build off these works by using transfer learning to pretrain BERT on larger domain specific corpora. In comparison to the finetuning corpus used in [9], our Parler dataset is nearly 30 times the size and our 4chan dataset is comparable in size. However, the issue of accurate human labeling of hate speech data remains a problem.

We hypothesize with the related works that improvements in hate speech classification can be made by using transfer learning from domain specific pretraining on large corpora of text. Independently, we hypothesize that pretraining on corpora of a similar domain to the finetuning dataset will improve. This would require a smaller pretraining corpora but limit the models ability to classify hate speech generally. We test both hypotheses in this paper. Aside from our baseline, we separately pretrain BERT on corpora from Parler and 4chan and implement exBERT, a BERT variant specifically built for domain specific pretraining explained in section 4. We test our models on hate speech datasets from Twitter [1], Reddit [2], and Gab [2], explained further in section 5.1. We observed similar results to [9], as our models did not surpass the baseline in multiclass classification in the Twitter (Davidson et al.) dataset. However, we observed improvements in our Parler and 4chan models on the Reddit and Gab datasets, respectively. This may be due to the size of the training corpora, the similarity of the testing domains, or improvements in the labeling of these latter finetuning datasets.

3 Related Work

3.1 Classification of Hate Speech

Hate speech classification has been a topic of interest since the advent of social media. However, classifying online hate speech is a challenging task, as spelling and usage of online hate speech and slurs change dramatically and rapidly [10]. While works like Waseem et al. [11] aim to create a typology of online hate speech, there does not exist a systematic way to classify hate speech, nor is there a shared benchmark dataset for such tasks. Davidson et al. [1] created a widely used multiclass hate speech dataset, used in this paper, and illustrates the difficulty classifiers face when attempting to distinguish hateful from offensive language or racist from sexist language. The granularity of this task and the imperfect and subjective process of human labeling the data correctly is explored in this and other works. Authors have found inconsistencies and mislabeled data in the datasets of Waseem et al. [11] and Davidson et al. [1]. Grondahl et al. [12] found that improving the accuracy of data labeling is equally if not more important than the choice of model architecture when it comes to model improvements.

Grondahl et al. [12] tested architectures including LSTM, CNN, CNN+GRU, SVM, and seq2seq on hate speech classification and observed that the CNN+GRU models performed the best but concluded that none of the models were effective at capturing the understanding of hate speech generally and performed poorly across datasets. Rizoiu et al. [4] saw improvements by using newer model architecture like bidirectional LSTMs and pretrained ELMo word embeddings to apply transfer learning of the embeddings from hate speech corpora. The use of pretrained word embeddings and transfer learning in Rizoiu et al. [4] improved the model's ability to create embeddings specific to the task at hand while also embedding the general meaning of hate speech. These related works noted that their models struggled in its ability to classify hate speech in context. The development of transformer models like BERT improved on this limitation and resulted in state-of-the-art results as shown in Mozafari et al. [8], inspiring the use of BERT in our paper. This work tested various combinations of neural network architecture including BERT base, BERT+Nonlinear layers, BERT+LSTM, and BERT+CNN. They observed that BERT+CNN performed the best. They believe that BERT's ability to embed the meaning of hate speech in context due to the finetuning of the embeddings using transfer learning. This work notes incorrect data labeling and training on larger corpora as a means of future improvement. We build off of this suggestion by pretraining BERT on larger pretraining corpora.

3.2 Domain Specific Models

The original BERT paper by Devlin et al. [13] suggests that additional pretraining on domain specific language should improve BERT's performance in other tasks including hate speech classification after pretraining on a hate speech corpus. However, pretraining BERT models on fringe corpora to improve hate speech classification is relatively unexplored and faces challenges including the rapidly changing and distinct vernaculars of online fringe communities. Works like de Gilbert et al. [14] have created domain specific corpora from white nationalist websites (`stormfront.org`) in order to improve domain specific pretraining efforts. Alatawi et al. [15] used the hate speech corpora created by de Gilbert et al. [14] and observed marginal improvements over the state-of-the-art BERT

results. Similarly, Isaken et al. [9] found that pretraining on domain specific corpora obtained equal or worse results to the state-of-the-art BERT results. Both of these works hypothesized that the limitations in the pretraining came from the size and quality of the training corpus. Other research into domain specific pretraining focuses on biomedical classification tasks. Models like SciBERT and BioBERT are two of the most successful examples. SciBERT builds a pretrained model from scratch with a domain specific vocabulary by looking at high frequency words and sub-words in scientific papers. This requires a large dataset and plenty of computing power as it does not use pretrained word embeddings. On the other hand, BioBERT builds off BERT's pretrained word embeddings before pretraining on a new corpus to modify the word embeddings. This model requires less data and computing power than SciBERT approach but is a less specialized model. Without expanding its vocabulary, it learns the biomedical specific material through combinations of sub-words that exist in the original BERT vocabulary. The exBERT [3] model builds off of the research from BioBERT and SciBERT by finding a more resource and computationally efficient way to train domain specific models without losing general model improvements. While exBERT is built for the biomedical domain, we explore modifying exBERT to classify hate speech using domain specific corpora and whether the domain of pretraining corpora can lead to classification improvements.

4 Approach

To understand how domain specific pretraining affects hate speech classification, we compare the classification performance of multiple models across three hate speech datasets. We use BERT base uncased as our baseline model. BERT base has 12 encoder layers, 768 hidden units, 12 attention heads, and 110 million parameters, and was trained on Wikipedia and BookCorpus (Figure 1). For our three other models we separately pretrain the BERT base uncased model on 4chan and Parler data. We choose to split the Parler model into two while using the same corpus. One model is trained on one epoch and the other three epochs. We do this to investigate whether additional training improved classification. We used an Azure NCv3 virtual machine to pretrain our models. Throughout our paper we will refer to the baseline BERT base uncased model as *baseline*, the model pretrained on 4chan data as *4chan*, the model pretrained on Parler data for one epoch as $Parler_1$, and the model pretrained on Parler data for three epochs as $Parler_3$. For our last model we implemented exBERT, [3] a technique to train BERT on a domain specific vocabulary. We will refer to this model as the exBERT model. This method creates an embedding layer to be optimized during pretraining with an extension vocabulary (Figure 2). The pretraining extension module creates new embeddings for a domain specific corpus separately from the original pretrained vocabulary layer. exBERT uses these embeddings to augment the original embeddings to make them more domain specific (Figure 2). The exBERT layers ($T_{ext}(.)$) augment each layer of the original BERT model ($T_{ofs}(.)$) using the following equation to compute the output of the hidden layer.

$$H^{l+1} = T_{ofs}(H^l) * \sigma(w(H^l)) + T_{ext}(H^l) * (1 - \sigma(w(H^l)) \quad (1)$$

Here, H^{l+1} is the output of the current hidden layer, H^l is the output of the previous layer, and w is fully connected layer that outputs the weights for the weighted summation of the two embedding layers. After we obtained weights for the pretrained models, we finetuned the models on each of the three hate speech datasets. The BERT language model learns word embeddings to attempt to understand what a word means in context. Only the vectors from the final encoder are relevant to the classification layer. The finetune layer is a feed forward neural network with one input layer, one hidden layer, one output layer, one softmax layer, and dropout.

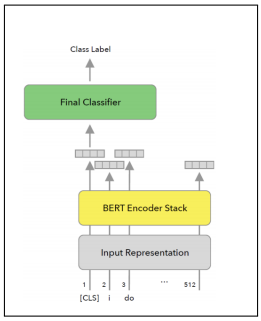

Figure 1: BERT Architecture [9]

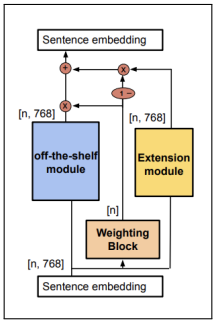

Figure 2: exBERT Architecture [3]

5 Experiments

5.1 Data

We use two unlabeled corpora for domain specific pretraining and three labeled hate speech corpora on which we finetune and evaluate our models, summarized in Table 5. The task for all finetuning datasets is to perform hate speech classification. The datasets are either multiclass (hateful, offensive, neither) or binary (hateful, non-hateful), explained below. For finetuning datasets, we used a random 80-10-10 train-val-test split using Pytorch's random dataset splitting function, and modify our models' word embeddings for the given domain through transfer learning. Note that hate speech is not well defined and though all three datasets are labeled by humans, determining whether a phrase is hate speech or not is subjective and may not accurately represent hateful language, affecting downstream classifications tasks. Datasets are refered to by the bolded names below.

5.1.1 Finetuning dataset from Davidson et. al. (2017)

The first dataset was collected by **Davidson** et. al. [16] and consists of 24,802 labeled tweets obtained from a hate speech lexicon (`hatebase.org`) and labeled by CrowdFlower (CF) workers. CF workers were asked to label each tweet as one of three categories: "hateful", "offensive", or "neither". Here, hateful speech is defined as abusive or threatening speech that targets disadvantaged social groups in a potentially harmful manner. Hate speech is generally directed at groups on the basis of race, gender, or sexual orientation. Offensive speech may contain vulgar words but not be directed at a disadvantaged group. The data consists of 1,431 (5.77%) cases of hate speech, 19,204 (77.43%) cases of offensive language, and 4,166 (16.80%) cases of neither hate speech nor offensive language.

5.1.2 Finetuning datasets from Qian et. al. (2019)

These datasets were constructed by Qian et. al. [2] from posts on Reddit and Gab. These datasets contain similar popular hate keywords but the distributions of said keywords differ greatly, illustrating that the two datasets are derived from different domains of the internet. Both datasets are split into the categories "Hateful" and "Not hateful" and were labeled by hand. **Reddit:** This dataset consists of 5,020 posts, including 22,324 comments, from toxic subreddits including *r/The_Donald, r/MensRights, and r/Imgoingtohellforthis*, among others. The authors collected the 200 most popular posts (with 4.5 comments on average) for each subbreddit and used hate keywords from [17] to identify potentially hateful comments. 3,847 (76.6%) of the conversations contained hate speech. When breaking the conversations into comments there are 5,257 (23.6%) labeled as hate speech and 17,067 (76.4%) labeled as non-hate speech. **Gab:** This dataset was constructed from all posts on the platform in October 2018. It contains 11,825 conversations containing 33,776 comments (3 comments per post on average). 94.5% of these conversations contain hate speech and 14,614 (43.3%) comments are labeled as hate speech and 19,162 (56.7%) are labeled as non-hate speech.

5.1.3 Pretraining data

Example text from these datasets are included in Tables 6 and 7. We chose to pretrain our models on 4chan and Parler due to the high amounts of hateful language found on these sites. They are both presumed to occupy a similar, yet distinct, domains on the internet as Reddit and Gab and serve to test whether pretraining on similar domains improves classification. **4chan:** We manually scraped over 50,000 posts from 4chan's /pol/ channel on Nov. 3, 2020. The /pol/ channel is notorious for its toxic speech. **Parler:** We used the Parler dataset collected by [18] containing 183 million posts. For pretraining, we used a subset of 4.7 million posts.

5.2 Evaluation method

We are using precision, recall, F1 scores, accuracy, and matthews clustering coefficient (MCC) [19] as our evaluation metrics. These metrics provide insights into the misclassifications of the models and are used by related works. We use precision, recall, F1, and accuracy to measure and interpret the number of correctly and incorrectly labeled tweets and to compare to the results of previous works. We chose to use the MCC following [1] as it provides a metric of comparison for classification tasks when the class sizes are imbalanced. This is helpful for evaluating our models on the Davidson and Reddit datasets which are quite imbalanced. The MCC is a correlation coefficient between -1 and 1 where a score of 1 means perfect prediction.

5.3 Experimental details

For the baseline model we used the off the shelf BERT model [20]. For the three pretrained models (4chan, Parler_1, and Parler_3) we started from BERT uncased, and then continued training on the 4chan and Parler datasets. We used an AdamW optimizer with a learning rate of $2e^{-5}$ and an epsilon of $1e^{-8}$. For training we used a batch size of 16. We trained our model for 3 epochs on 4chan, 1

epoch on Parler_1 and 3 epochs on Parler_3. We trained on text files where each line was an individual post and tokenized using the prebuilt BERT tokenizer, which takes token sequences of maximum length 512 and lowercases all of them. For all models we used a masking rate of 0.15, meaning 15% of tokens in a sentence are masked. We trained our exBERT model with an extended vocabulary and an extension module. We used the same 4chan pretraining corpus and the word embeddings of the baseline, and trained the exBERT model for 2 epochs with a learning rate of $1e^{-4}$ and a dropout probability of 0.1. The model uses GELU as the activation function. When finetuning the first four models (baseline, 4chan, Parler_1, Parler_3), we used the HuggingFace [20] classifier with our specific pretrained weights. We designed our datasets, data preprocessing, classifier finetuning, and evaluation system to be compatible with the HuggingFace [20] architecture. The classifier uses the HuggingFace Adam optimizer with weight decay fix, warmup, and linear learning rate decay. For training we used a batch size of 16 and ran our model for 4 epochs. For finetuning exBERT we used a batch size of 8 over 8 epochs with a maximum sequence length of 512 to train the classifiers. To validate that our domain specific pretraining worked, we loaded the weights into the HuggingFace MaskedLM model and performed text generation (Table 8). The COVID-19 references convince us that our pretraining worked. Given the ethical concerns regarding hate speech text generation, we only used this as a validation check.

5.4 Results

Model		Baseline	4chan	Parler_1	Parler_3	exBERT
Not hate	P	0.910	0.928	0.934	**0.938**	0.632
	R	0.890	**0.908**	0.871	0.880	**1.000***
	F_1	0.900	**0.918**	0.901	0.908	0.775
Hate	P	0.791	**0.834**	0.775	0.783	0.000
	R	0.825	0.868	0.878	**0.881**	0.000
	F_1	0.808	**0.851**	0.823	0.829	0.000
	M	0.698	0.719	0.729	**0.740**	—
	A	0.869	**0.894**	0.873	0.880	0.632
Macro avg.	F_1	0.854	**0.884**	0.862	0.868	0.387

Table 1: Result from model experiments on our Gab dataset. *See section 6.1

Model		Baseline	4chan	Parler_1	Parler_3	exBERT
Not hate	P	0.916	**0.944**	0.938	0.937	0.805
	R	0.932	0.937	0.934	**0.941**	**1.000***
	F_1	0.924	**0.941**	0.936	0.939	0.892
Hate	P	0.688	0.660	0.672	**0.709**	0.000
	R	0.639	0.689	0.688	**0.695**	0.000
	F_1	0.663	0.675	0.680	**0.702**	0.000
	M	0.577	0.587	0.649	**0.651**	—
	A	0.876	**0.899**	0.893	**0.899**	0.805
Macro avg.	F_1	0.793	0.808	0.808	**0.820**	0.446

Table 2: Result from model experiments on our Reddit dataset. *See section 6.1

Model		Baseline	4chan	Parler_1	Parler_3	exBERT
Hateful	P	0.500	0.455	0.504	**0.518**	0.000
	R	**0.449**	0.437	0.420	0.390	0.000
	F_1	**0.473**	0.445	0.458	0.445	0.000
Offensive	P	**0.948**	0.945	0.941	0.936	0.772
	R	0.954	0.952	0.956	**0.959**	**0.990***
	F_1	**0.951**	0.948	0.949	0.947	0.867
Neither	P	0.889	0.890	0.873	**0.901**	0.000
	R	**0.887**	0.870	0.855	0.876	0.000
	F_1	**0.888**	0.880	0.863	**0.888**	0.000
	M	**0.762**	0.747	0.751	0.754	—
	A	**0.918**	0.912	0.909	0.911	0.766
Macro avg.	F_1	**0.771**	0.758	0.757	0.760	0.289

Table 3: Result from model experiments on our Davidson dataset. *See section 6.1

In Tables 1, 2, and 3, P represents Precision, R represents Recall, F_1 represents the F_1 score, M represents the Matthew's clustering coefficient, and A represents accuracy. Each class is represented by a row where the last row is the model's overall performance. The results found in the table above are better than expected and indicate we are taking the correct approach to test our hypotheses. We notice that the Parler_3 model performs quite well on all datasets, which strengthens the assumption that improvement comes with larger pretraining corpora. The results for the Davidson dataset follow results from [9] where additional pretraining did not improve upon the baseline model. This may stem from the general difficulty of multiclass classification or the differences between the vernacular of hate speech found in Davidson and the pretraining datasets. Results from Gab and Reddit are better than expected as our pretrained models outperform the baseline. We believe 4chan performs

better on the Gab dataset given their similar vocabularies, allowing it to pick up on some of Gab's coded hate speech. Our Parler$_3$ model may perform better on the Reddit dataset due to the size of the pretraining corpus, allowing it to pick up on hate speech in context better. It may also perform better on Reddit than Gab given Reddit is more mainstream platform and likely contain less coded language. Our results are not surprising for exBERT as the lack of documentation and highly buggy codebase made the finetuning method relatively incompatible with these datasets despite weeks of debugging.

6 Analysis

6.1 Error Analysis

The results show that it is difficult for hate speech classifiers to separate hateful language from offensive language, as seen in Figures 3 and 4. Confusion matrices for all of the models can be found in Appendix B. Below we compare the baseline and 4chan models which had the best and worst results, respectively, for the Davidson dataset. Both models predicted similar amounts of hate speech as offensive language. However, the baseline model was much better at predicting offensive language, unlike the 4chan model. This may be due in part to the hateful corpus the 4chan model was trained on, leading it to classify a larger number of observations as hateful. [1] and [9] achieved the same results for this dataset. This may be due to the imbalanced dataset; there are an overwhelming number of offensive labeled data points and thus the model was not trained on enough examples of hate speech.

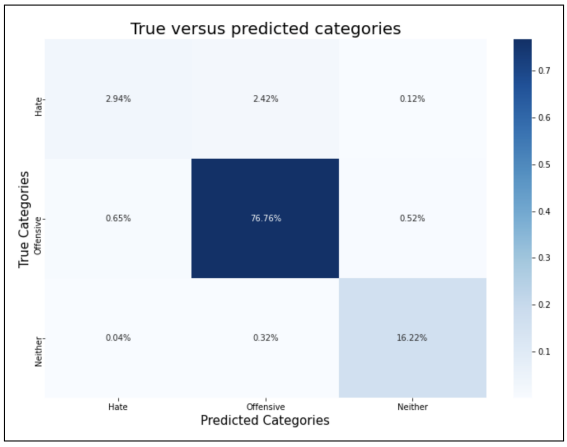

Figure 3: Results of vanilla BERT on Davidson dataset

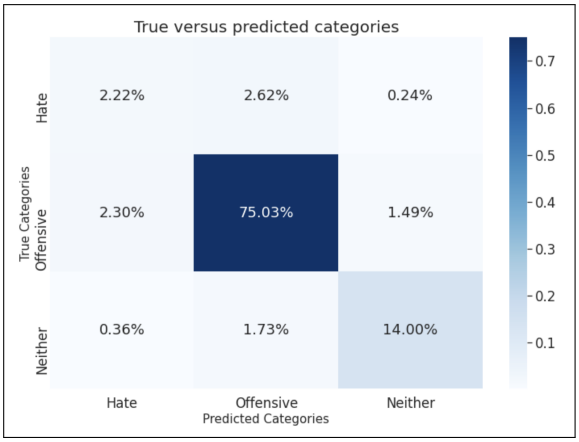

Figure 4: Results of BERT finetuned on 4chan data on Davidson dataset

When analyzing the Gab and Reddit datasets in Tables 1 and 2, we notice that the 4chan model and the Parler$_3$ model perform the best, respectively, based on the accuracy metric. Parler$_3$ is significantly better at classifying hateful speech in the Reddit dataset than all other models as illustrated in confusion matrix 8 as well as by its "Hate" F_1 score. The confusion matrices also illustrate how Parler$_3$ and 4chan both incorrectly predicted roughly 5% of the Reddit data as non-hateful when in fact it was hateful. This may be due to the imbalance of the dataset which has nearly 3 times the amount of non-hateful examples. As we examine below, these examples may have contained a hateful word in a non-hateful context which may have confused the classifiers.

The 4chan model outperforms all models at classifying both hateful and non-hateful speech in the Gab dataset as illustrated by its "Not Hate" F_1 score and confusion matrix 10. 4chan's high precision score in Gab tells us it overpredicted on non-hateful speech. Given 4chan's hate-skewed predictions in other datasets and the hateful corpus it was trained on, we infer that 4chan was able to capture and embed hate speech in context during transfer learning. This result tells us that 4chan has a good understanding of the Gab vocabulary and hate speech patterns, strengthening the idea that training on similar domains lead to classification improvements. Selected examples examined below further strengthen this belief. We also notice that all models have higher precision and recall scores across both classes for the Gab dataset when compared to both Reddit and Davidson. This may be due the balance between classes in the dataset. We believe that the Parler$_3$ model performs well on the Gab dataset due to the size of the corpus it was trained on, helping it embed the context of hate speech. Finally, the exBERT model does not perform well compared to the other models aside from achieving perfect or near perfect recall in one category of each dataset. However, we are skeptical of all results

from this model due to reasons explained in section 6.3. We decided to include these results as a baseline, albeit a broken baseline, to observe future improvements.

6.2 Commenting on Selected examples

Below we examine some correctly and incorrectly classified instances across all of the models. We chose to focus on the baseline, 4chan, and Parler_3 as these models all performed either the best or the worst on one of the testing datasets. The key insight as to why our 4chan and Parler pretrained models outperform the baseline on Reddit and Gab stems from the vocab files and tokenizers generated from the pretrain. As we can see Table 4, all of the hate-speech examples have key words that identify them as being highly likely candiates for hate-speech. The BERT base vocab file has very few slurs in it, and breaks down words like the n-word, and f***got into multiple subwords. In contrast, the 4chan and Parler vocab files have full word emeddings for slurs, and importantly, often have word embeddings for mispelled slurs. This may have helped in the second row of Table 4 where the word "Muslim" is misspelled yet our 4chan model was able to capture the anti-Islamic sentiment. In addition to typos, given the informality of our online speech, abbreviations for slurs are quite common and are picked up by our 4chan and Parler models. Further, it is unsurprising that the baseline classifier missed many anti-Semitic and anti-Muslim sentences given that the words "jew," "jewish," and "muslim" may appear in many non-hateful contexts on the corpus BERT was trained on whereas they may appear much more frequently in a hate speech context on 4chan or Parler.

Model Classification				
Correct	**Incorrect**	**Dataset**	**Example**	**Correct class**
4chan	Baseline	Gab	*a vile f*g**t wearing tight pants*	Hate
4chan	Baseline	Gab	*you decided you just wanted to swim with the shit sucking moozlum royals of britanistan*	Hate
4chan	Baseline	Gab	*the magatards have turned an anti white nut job into a national hero. even our guy tucker was waxing poetic about kanye. the jews probably laugh their asses off.*	Hate
Parler_3	Baseline	Reddit	*what a bunch of chundering c*nt bags*	Hate
Parler_3	Baseline	Reddit	*so you want free speech for you but not for anyone else? whatever you k*ke f**ktard go gas yourself*	Hate
Baseline	Parler_3	Davidson	*ni**a69 true dat ni**	Offensive
Baseline	Parler_3	Davidson	*jihadist joe: we muslims have no military honour whatsoever we are sub human savages that slaughter unarmed men women amp; children*	Hate
Baseline	Parler_3	Davidson	*a bird just flew into my window?????? i hope he's okay*	Neither

Table 4: Examples misclassified by various models on all datasets.

After investigating datasets where our models outperformed the baselines, we were then curious to see why on the Davidson classification, the more we trained, the worse we performed. From our inspection of the Davidson dataset, the hate speech class seems to include fewer slurs which are prevalent in our pretraining corpora. 4chan, Parler, and Gab are for the most part unmoderated communities unlike Twitter, which Davidson draws its corpus from, making the distribution of its vocabulary drastically different from the pretraining corpora. Additionally, Reddit and Gab datasets were constructed by finding hate speech through key word searches which would make slur embedding incredibly important for accurate classification of these datasets. This may be one reason our pretrained models perform poorly on Davidson and better on Reddit and Gab. Table 4 shows examples from each Davidson class where the base model predicted correctly, and our Parler_3 model missclassified. Davidson is our only muticlass dataset, which means the models need to differentiate between offensive and hate speech. As shown in the final row (row 8), Parler_3 classified a mundane phrase as offensive. The tendency of this model to overclassify on hate speech may be due in part to the more hateful nature of the Parler dataset which may impact its word embeddings even for mundane words. In row 6 of Table 4, we see an example where Parler_3 misclassifies a phrase as hate speech, while the baseline correct predicts it as offensive. It is possible our Parler model recognizes the n-word as hateful and classifies it as such despite the word appearing in a non-hateful context. This example also illustrated the difficulty of labeling data even for humans as this phrase may be labeled differently depending on the definition of offensive language. In row 7 we observe hateful speech that does not contain a slur. The fact that Parler_3 is unable to classify it as such despite its clearly hateful nature supports the argument that it is basing much of its classification on the presence of slurs. We also investigated the misclassifications of the baseline and Parler_1 models on the Gab dataset included in Table 9 and further misclassifications in Table 10.

6.3 exBERT

We did not observe improvements with exBERT for the following reasons. First, the extension vocabularies likely changed the original word and subword embeddings due to crossover with the original BERT vocabulary. This may have led our model to lose BERT's general understanding of the meaning of certain words. If the embedding relies too heavily on what it learned from the domain specific text, it may not be able to accurately interpret word meaning and instead embeds a narrow word definition. The architecture of exBERT is best suited for vocabulary expansion and may work best for domains with little vocabulary overlap. In the original paper [3], 56% of the vocabulary in the domain specific corpus does not exist in the original BERT vocabulary unlike our pretraining corpora where only 37.58% of the 4chan vocabulary was new to BERT. Due to the importance of BERT's pretrained embeddings, overlapping vocabulary may alter word embeddings such that crucial word understanding is lost and made too domain specific. Finally, the exBERT codebase required significant restructuring to both train and finetune. Due to the disorganization and lack of documentation, it is reasonable to believe that by refactoring the code to make it usable, some of the architectural choices implemented by the original authors may not have been reflected in the pretraining we were able to do. We were unable to pretrain exBERT on our Parler dataset and unable to obtain significant results after pretraining on 4chan as the model as it would only predict on a subset of the hate classes. However, we were able to adapt the code to finetune the model on COVID-19 misinformation classification (see Appendix C).

6.4 Limitations

One limitation is the quality of the labeling in hate speech datasets. Without a universal definition, hate speech can be labeled subjectively; some may disagree with others about what constitutes offensive vs. hate speech. On top of subjectivity, there is also the issue of mislabeling. Given these limitations, it is possible there is bias in how these datasets are labeled and therefore our results may not accurately classify offensive and hate speech accurately. Our Parler_3 model, which showed some of the largest improvements, trained for over 100 hours on 5 millions posts on an Azure NCv3 server and improved drastically with more training time. This is limiting as such corpora are not always available and it is also both computationally expensive and environmentally wasteful. Further, despite improvements seen by pretraining on a smaller but domain similar 4chan dataset may limit the model by improving its ability to classify hate speech within a specific domain but fail in general, as we saw during its evaluation on the Davidson dataset. Another limitation is the nature of the corpora on which our models are pretrained. The Parler and 4chan posts used for pretraining both contain high volume of slurs and therefore may have a harder time identifying hate speech that does not include slurs or other informal hateful words. This may also lead these models to incorrectly classify a phrase as hate speech should it notice the slur even without a hateful context.

7 Conclusions and Future Work

We aimed to test whether pretraining on large and domain specific corpora improved classification. To do so we pretrained BERT base on domain specific corpora and compared it to our baseline model on a number of hate speech classification tasks. We observed improvements over the baseline by our Parler_3 and 4chan models in Reddit and Gab datasets, respectively. Parler_3's large pretraining corpora and 4chan's domain similarity to Gab are likely reasons for the improvement, following from our hypotheses. Further, we observed improvements across all metrics (accuracy, precision, recall, F1, MCC) for our pretrained models when classifying on Reddit and Gab, potentially indicating better data labeling or improvements due to pretraining in similar domains. For the Davidson dataset, we saw a decrease in performance for pretrained models, even with additional training. This may be because the makeup of our pretrained corpora lead these models to rely on slurs and informal language to detect hate speech. Our exBERT model did not have metrics with which we were able to compare to the rest of our results.

These findings are tempered by the quality of the data and rely on the accuracy of the labeling of said data, which is one avenue for future improvements. There are many potential avenues for future work, but one that specifically excites us is a continuation of our COVID misinformation work. Our Parler_3 model observed the same accuracy as the baseline, but had far worse precision and far greater recall. This suggests both the baseline and the Parler_3 model are correctly classifying examples that the other failed to do. We are also hoping to further explore the idea of finding a balance between using smaller datasets that are able to capture general hate speech in context. We saw the benefits of training on a large corpus as well as training on domains similar to the evaluation domain but finding a middle ground between these two approaches to increase performance would improve hate speech classification and make it more accessible for those with fewer resources.

References

[1] Thomas Davidson et. al. Automated hate speech detection and the problem of offensive language. In *AAAI Conference on Web and Social Media*, ICWSM '17, 2017.

[2] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech, 2019.

[3] Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online, November 2020. Association for Computational Linguistics.

[4] Marian-Andrei Rizoiu, Tianyu Wang, Gabriela Ferraro, and Hanna Suominen. Transfer learning for hate speech detection in social media, 2019.

[5] Drew Harwell. Three mass shootings this year began with a hateful screed on 8chan. its founder calls it a terrorist refuge in plain sight., Aug 2019.

[6] Freddy Cruz and Hannah Gais. Far-right insurrectionists organized capitol siege on parler, Jan 2021.

[7] Quang Huu Pham, Viet Anh Nguyen, Linh Bao Doan, Ngoc N. Tran, and Ta Minh Thanh. From universal language model to downstream task: Improving roberta-based vietnamese hate speech detection. *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, Nov 2020.

[8] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A bert-based transfer learning approach for hate speech detection in online social media, 2019.

[9] Vebjørn Isaksen and Björn Gambäck. Using transfer-based language models to detect hateful and offensive language online. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 16–27, Online, November 2020. Association for Computational Linguistics.

[10] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[11] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks, 2017.

[12] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, page 2–12, New York, NY, USA, 2018. Association for Computing Machinery.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[14] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[15] Hind Saleh Alatawi, Areej Maatog Alhothali, and Kawthar Mustafa Moria. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert, 2020.

[16] Available at `https://github.com/t-davidson/hate-speech-and-offensive-language`.

[17] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. Peer to peer hate: Hate speech instigators and their targets, 2018.

[18] Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. An early look at the parler online social network, 01 2021.

[19] Giuseppe Jurman et. al. A comparison of MCC and CEN Error Measures in Multi-Class Prediction. 2012.

[20] Hugging Face. Transformers. 2020. Available at `https://github.com/huggingface/transformers/tree/master/examples/language-modeling`.

A Tables

	Task	Source	# posts	Categories
Davidson et. al. (2017)	Testing	Twitter	25k	hateful, offensive, neither
Qian et. al. (2019)	Testing	Reddit	22k	hateful, not hateful
Qian et. al. (2019)	Testing	Gab	34k	hateful, not hateful
Aliapoulios et. al. (2021)	Finetuning	Parler	4.7M	n/a
Manually collected	Finetuning	4chan	100k	n/a

Table 5: Summary of all datasets (pretraining and finetuning) used in experiments.

ID	Example Post
1	NAZI PIGS disguised as Police escorted 4 Buses of Antifa into the Capitol
2	!!!BE PREPARED TO DEAL WITH BLADES AND OTHER SHARP EDGE WEAPONS!!!
3	A Jew just drove by talking sh** and one of the things he said was that he will kill me ...

Table 6: Example posts from our Parler dataset containing 185 million Parler posts.

ID	Example Post
1	wow for real? enjoy jail fa**ot.
2	Why cant western f***ers just learn to handle hard times instead of fleeing like rats
3	Full of Ch****. No meat available for the average people

Table 7: Example posts from our 4chan dataset. All posts occurred on the /pol/ channel between 8am and 12am on the day of Nov. 3, 2020.

Prompt	Generated Text
Fauci is	"a traitor to america now too"
Masks are	"illegal"
Voting	"machines down in ga today"

Table 8: Text generation using the weights of our BERT model pretrained on Parler data to validate pretraining efforts.

Model	Dataset	Example	Correct Class
$Parler_1$	Gab	*evading a perm ban that anon calm fgt is reporting us all*	Hate
$Parler_1$	Gab	*yuck i hate that cu*t anglin she looks like a person we defend.*	Hate
$Parler_1$	Gab	*all this group is doing is dividing whites. whites dividing whites how very jewish, applaud yourselves*	Not Hate
Baseline	Gab	*bring it you goat f***ing f**got*	Not Hate
Baseline	Gab	*they rarely do anything about wetb***s because ignoring the crime is easier than having to perpetually raid the businesses that hire illegals*	Hate
Baseline	Gab	*keep and open mind and listen. please feel free to speak freely however the jew does not censor you here*	Hate

Table 9: Examples misclassified by the $Parler_1$ and and baseline models on Gab dataset.

Model Classification				
Correct	**Incorrect**	**Dataset**	**Example**	**Correct class**
4chan	Baseline	Gab	*a vile f**got wearing tight pants*	Hate
4chan	Baseline	Gab	*you decided you just wanted to swim with the shit sucking moozlum royals of britanistan*	Hate
4chan	Baseline	Gab	*the magatards have turned an anti white nut job into a national hero. even our guy tucker was waxing poetic about kanye. the jews probably laugh their asses off.*	Hate
4chan	Baseline	Gab	*We just need to find out who the good jews are to figure out who doesnt want to destroy us. then we can destroy ourselves without breeding because #maga #trump*	Hate
$Parler_3$	Baseline	Reddit	*can we make this mad c*nt the gold standard for gayness?*	Hate
$Parler_3$	Baseline	Reddit	*what a bunch of chundering c*nt bags*	Hate
$Parler_3$	Baseline	Reddit	*so you want free speech for you but not for anyone else? whatever you k*ke f**ktard go gas yourself*	Hate
Baseline	$Parler_3$	Davidson	*ni**a69 true dat ni**	Offensive
Baseline	$Parler_3$	Davidson	*jihadist joe: we muslims have no military honour whatsoever we are sub human savages that slaughter unarmed men women amp; children*	Hate
Baseline	$Parler_3$	Davidson	*a bird just flew into my window?????? i hope he's okay*	Neither

Table 10: Examples misclassified by various models on all datasets.

B Confusion Matrices

B.1 Davidson Dataset

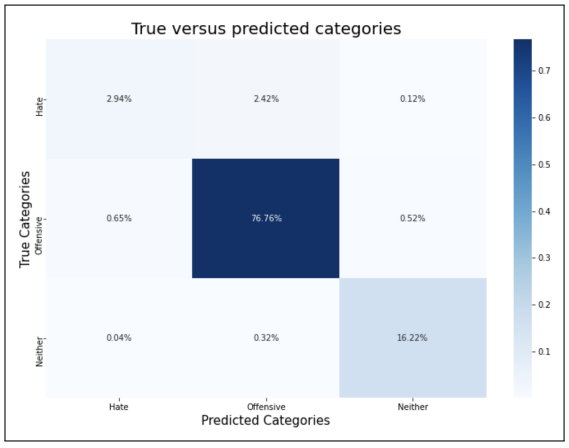

Figure 5: Results of vanilla BERT on Davidson dataset

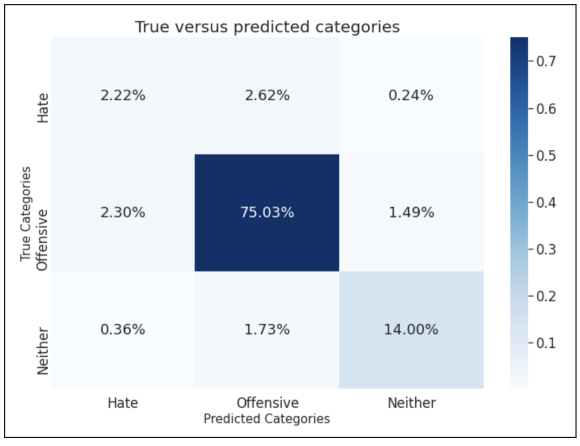

Figure 6: Results of BERT finetuned on 4chan data on Davidson dataset

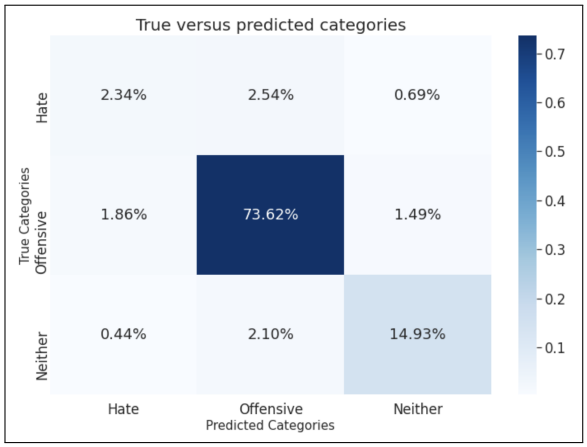

Figure 7: Results of BERT finetuned on 1 epoch of Parler data on Davidson dataset

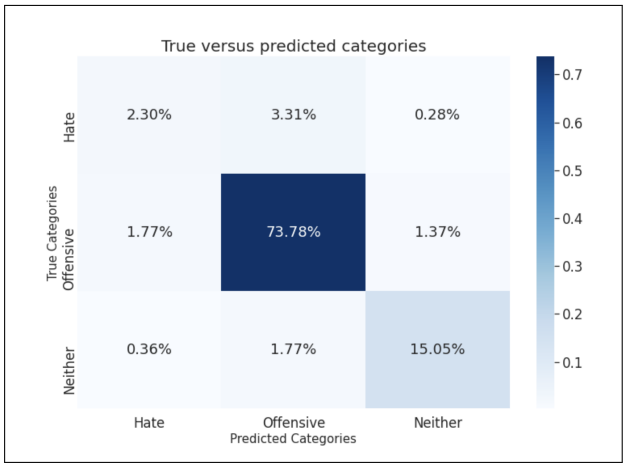

Figure 8: Results of BERT finetuned on 3 epochs of Parler data on Davidson dataset

B.2 Reddit Dataset

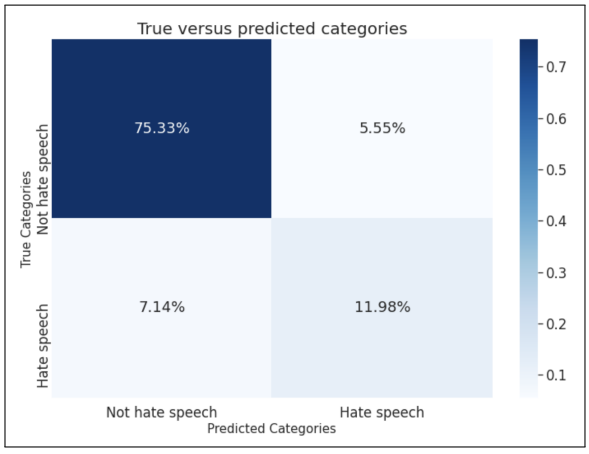

Figure 9: Results of vanilla BERT on Reddit dataset

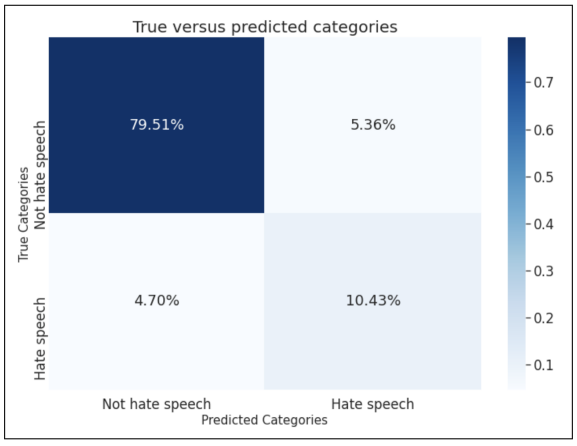

Figure 10: Results of BERT finetuned on 4chan data on Reddit dataset

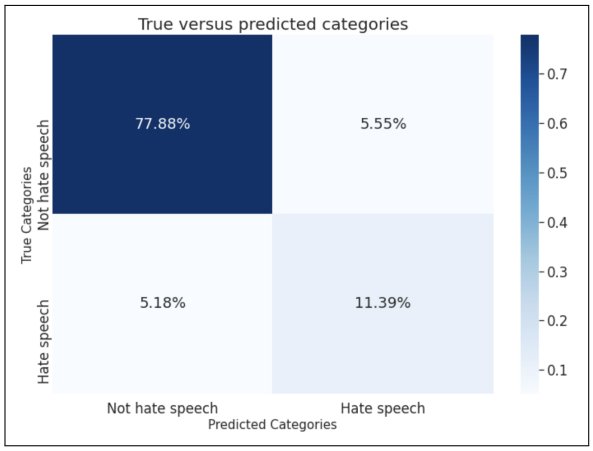

Figure 11: Results of BERT finetuned on 1 epoch of Parler data on Reddit dataset

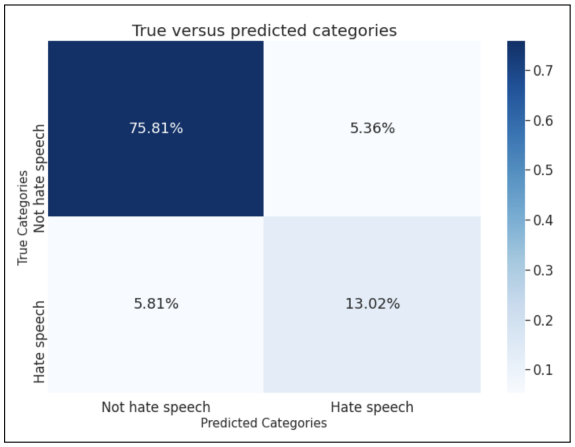

Figure 12: Results of BERT finetuned on 3 epochs of Parler data on Reddit dataset

B.3 Gab Dataset

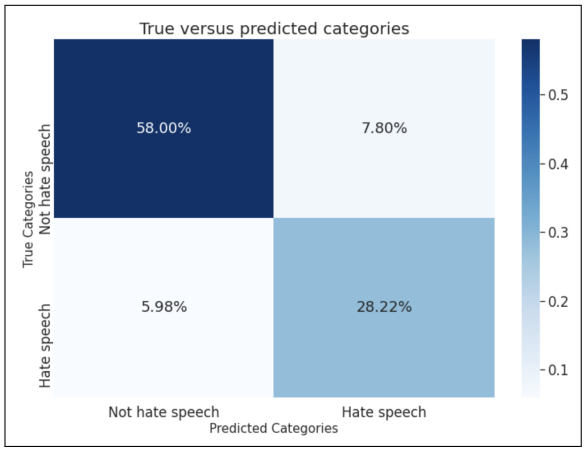

Figure 13: Results of vanilla BERT on Gab dataset

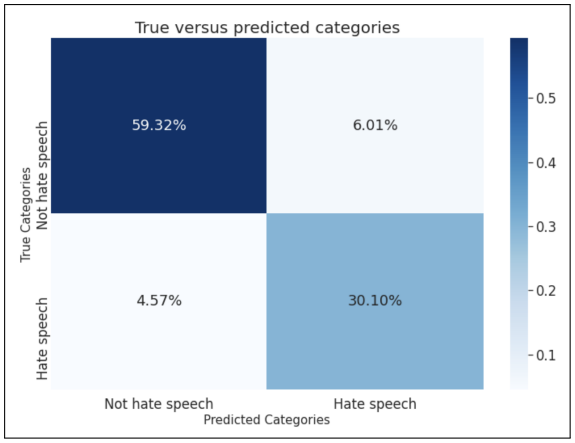

Figure 14: Results of BERT finetuned on 4chan data on Gab dataset

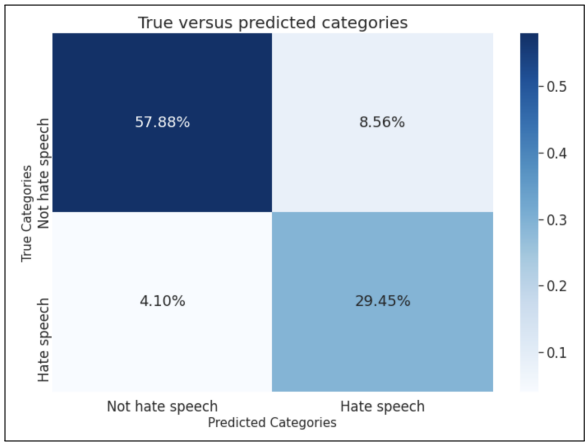

Figure 15: Results of BERT finetuned on 1 epoch of Parler data on Gab dataset

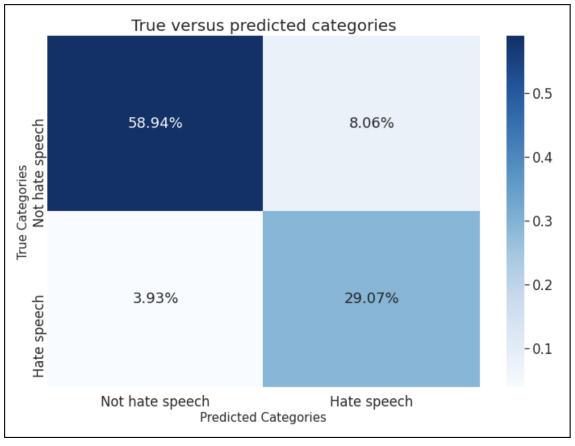

Figure 16: Results of BERT finetuned on 3 epochs of Parler data on Gab dataset

C COVID Misinformation Classification

Model		Baseline	4chan	Parler_1	Parler_3	exBERT
Fake	P	**0.905**	0.902	0.811	0.888	0.556
	R	0.741	0.754	0.741	**0.874**	0.098
Macro Avg	M	**0.746**	0.67	0.574	0.742	0.118
	A	**0.876**	0.828	0.786	0.872	0.673

Table 11: Result from model experiments on our covid misinformation dataset