The Dark Side of Language: Using NLP to Combat Hate Speech

Shalini Chawla, shalini_chawla@berkeley.edu

Adam Hyman, adamhyman@berkeley.edu

Sreeram Ravinoothala, sreeram@berkeley.edu

Abstract

The rise of social media has enabled individuals to spread hate anonymously, creating a greater danger than ever before because its impact is no longer limited to a specific social group or geographic region. It is very important for social media companies to accurately identify hateful content and a lot of work has already been done to classify online content as toxic vs non-toxic for content moderation. We would like to extend this classification from binary to multi-label, in order to identify the specific hate categories that each toxic content belongs to. This information can be used by socio-political studies exploring the relationship between specific social-political events and their impact on triggering hateful content in online media. In this paper, we fine-tune three pre-trained large language models for a multi-label classification task to categorize text comments as *toxic*, *severe_toxic*, *obscene*, *threat*, *insult* and *identity_hate*. To tackle the imbalance of available training data, we use back translation using multiple intermediary languages to augment training data for the minority classes.

1 Introduction

Our objective for this project is to classify hate content by the type of toxicity.

Social media brings people from all areas of society to shared spaces. While a wide-variety of view points is usually a good thing, there is a small amount of speech, that social media platforms consider to be hurtful to the platform and the community they're trying to build, hate speech. Hate speech can be targeted against an individual or group based on inherent characteristics, like race or gender. The offensive speech disrupts social peace on these platforms and upsets the users. The social media platforms expend considerable resources to remove hate speech, by allowing users to flag offensive content and with automated algorithms, so the derogatory comments can be removed and negative effects mitigated as soon as possible.

With the advances in machine learning, many social media platforms have made significant progress in controlling the proliferation of hate speech by making use of classification algorithms. However, due to the complex nature of speech, where the meaning of words change based on context and over time, there is still much progress to be made in this area.

The ability to build accurate models to correctly classify hateful content is a challenging task due to the limited availability of labeled data required for training. The training data needs to be labeled by human annotators. In addition to being a tedious task, it also necessitates exposing the annotators to disturbing content in order to label it. The public data sets available today are highly imbalanced and have very limited samples for some of the minority categories including *threat*, *obscene* and *severe_toxic* compared to the large amount of non-toxic samples. We follow two approaches to deal with the limitations of data availability.

1) Pre-trained large language models have made it feasible to achieve better accuracy in many NLP tasks where the amount of training data is limited, by providing word embeddings that have already learned characteristics of natural language. Using these pre-learned embeddings as the input to a classification model enables the model to learn to classify the content using much less training data than it would have needed to learn the same relationship from scratch.

We use three large language models: BERT, T5 and XLNet and fine-tune them for a multi label classification task using our data set that has been

labeled with the 6 toxicity classes.

2) Our source data set is highly imbalanced and has very limited data for three out of the six categories. We use a combination of approaches to balance our data set to help the model balance learning across all categories.

We have three minority categories: *threat*, *severe-toxic* and *identity_hate* that have very low representation in our training data set. We use the TOXIGEN (Hartvigsen et al., 2022) dataset to augment data for the *identity_hate* class. To augment data for the other minority classes, we use back translation to generate additional samples from the existing data.

2 Background

Cyber-bullying was first documented at the turn of the century when internet use was no longer restricted to academic institutions, large corporations and government, but became common place as home PCs could connect to the internet. A 2002 publication to address bullying in schools included "sending malicious e-mails or text messages on mobile phones" as a form of bullying (Haworth-Roberts, 2002).

Within only 6 years, 22% of high school students reported having been cyber-bullied at least once (Smith et al., 2008). The figure quickly doubled while researchers at MIT built the first models to detect cyber-bullying (Dinakar et al., 2011) utilizing TF-IDF, bigrams and lists of profane words.

It would be about a decade before the focus on Attention (Vaswani et al., 2017) and pre-trained large language models like BERT (Devlin et al., 2018) made it possible to create the sophisticated models to do accurate classification that we have today.

There has been a lot of research and previous work done with focus on toxic content classification and different ways have been explored to improve the accuracy of classification.

The unavailability of enough quality labeled data has been widely accepted as a limitation in the ability to create accurate models for toxic content classification. Rastogi et al. (2020) has explored generating synthetic data using Easy Data Augmentation and back translation to augment the training data and reported improved recall and F1 scores. Hartvigsen et al. (2022) have explored the complexity of identifying implicit *identity_hate* targeted at minority groups and used GPT-2 to generate additional data to augment the current available human labeled data sets. Another issue has been the inconsistency in labeling methodologies across the publicly available data sets limiting the generalization of trained models that was studied by Fortuna et al. (2020).

3 Data

3.1 Source Data Set

We used the dataset from the Jigsaw Toxic Comment Classification Challenge (Adams et al., 2017) held by Kaggle, that contains Wikipedia comments, that were labelled by human raters for various types of toxic content. The data set consist of 3 files that includes the labeled training data, test data and the test labels (that was made available after the competition concluded). We combined the test data and labels file by joining them on the unique comment identifiers. The training data set had 159,569 comments while the test data set had 63,972 comments. We converted the comments into lowercase and cleaned the contents by removing punctuation and special characters. Once processed both the training and test data consisted of the text comments and binary labels for the 6 categories:

- toxic
- severe_toxic
- obscene
- threat
- insult
- identity_hate

3.2 Data Balancing

Our source data is highly imbalanced. Out of the 6 labeled classes, 3 are majority classes and 3 are minority classes.

3.3 Augmenting Minority Class Data

We increased the training data for minority classes to bring it closer in count to the majority classes.

For *severe_toxic* and *threat*, we used back-translation with Google Translate API to supple-

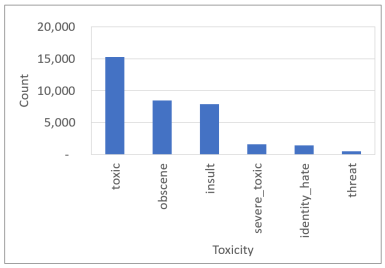

Figure 1: Toxicity Label Count

ment the training data. For the smallest class, *threat*, we back-translated all 476 comments using 3 languages - Chinese, Hindi and French - which along with the original comment quadrupled the number of comments in our training data. For *severe_toxic*, we back translated 500 comments using Chinese, to supplement our training data.

We supplemented *identity_hate* with 640 records created by Microsoft's Toxigen repository, which uses a large scale pre-trained language model and a classifier to generate synthetic hate speech. The repository includes two methods for generating hate speech: Demonstration-Based Prompting and an adversarial set up between a classifier and a generator. We included 320 records from each of the methods.

3.4 Undersampling Majority Class Data

To further balance the data set for the 6 identified classes, we selected all comments labelled with the minority classes, including augmented data. Since many of these comments are also labelled with the majority classes, we only needed to select a small amount of additional comments, that were labelled with the majority classes, but not the minority classes, in order to get a data set that was approximately balanced.

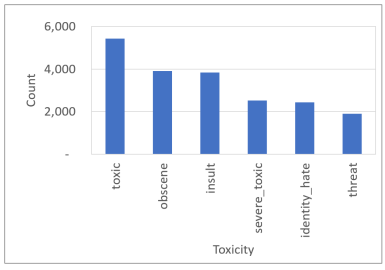

Figure 2: Label Count for Balanced Data

class	prec	recall	f1-score
toxic	0.62	0.40	0.49
severe_toxic	0.00	0.00	0.00
obscene	0.66	0.32	0.43
threat	0.00	0.00	0.00
insult	0.65	0.25	0.36
identity_hate	0.00	0.00	0.00
micro avg	0.64	0.31	0.42
macro avg	0.32	0.16	0.21
weighted avf	0.58	0.31	0.40
samples avg	0.04	0.03	0.03

Table 1: Baseline Model - Classification Report

4 Methods

4.1 Baseline Model - BERT fine-tuned on source data

Our baseline model is BERT base model (cased). We fine-tuned the model on the full training data set that was split 80:20 into training and validation and stratified on the 6 label classes. We added a hidden layer with 200 neurons and dropout layer with 0.3 before our final classification layer with 6 outputs to perform a multi-label classification. We trained the model for 2 epochs using learning rate of .00005 and batch size of 32.

The model performed very poorly for our 3 minority classes, completely failing to identify any of the *severe_toxic*, *threat* or *identity_hate* comments.

4.2 BERT fine-tuned with balanced data

Our baseline model's poor performance was due to the much smaller amount of training data for the minority classes. The model is favoring the majority classes and is learning to just return the label of the majority classes to maximize accuracy since the training data contains so few comments labeled with the minority classes. To help our model learn the features of the minority classes, we created a new balanced data set by undersampling our majority classes and augmenting the minority classes. We run the BERT model with the same hyperparameters as in our baseline with the balanced data set.

Balancing the dataset did not improve the classification score for minority classes, with none of the 3 minority classes being identified. Our data set size was reduced drastically after undersampling

but we still had 16,413 records which is an acceptable amount of data for fine-tuning BERT. Since BERT was originally trained on Wikipedia data, it has not seen enough toxic vocabulary, so is not able to learn enough from the training data to identify the minority classes.

4.3 Retraining BERT layers

Our next step was to systematically unfreeze the pre-trained BERT layers 2 at a time and retrain them using our data set. We used the original unbalanced data set and the balanced dataset we created and re-trained the top 2 BERT layers, followed by 4, 6 and all the layers. The first classifier with retraining only the top 2 BERT layers gave as a huge improvement in score for all three of the minority classes. As we retrained deeper layers, the total score kept moving upwards. For the full unbalanced dataset, the improvement stagnates after 4 layers. This has been observed in previous work done by Singh (2020). But we observed that with the smaller balanced dataset, the improvement continues as we retrain more layers with full retraining of all 12 layers giving us the best scores.

4.4 T5

We used T5-small (Raffel et al., 2020) on the unbalanced training dataset similar to BERT. The model was adopted from Jaunjale (2020). As part of the multi-label classification a language modeling head is used on top of the decoder. The model was run 2 epochs with 8 and 16 batch size. The low number of epochs was because of the time the model was taking to train.

The model did worse than BERT on every label except *severe_toxic*, which BERT had done poorly on. In addition, T5 failed to predict the *insult* majority class and two of the minority classes (*threat* and *identity_hate*) which can be seen from the lower ROC-AUC scores.

The same experiment was repeated with balanced dataset for 2 epochs and 16 batch size. The results were similar to what was seen with unbalanced dataset. It was still not predicting the *insult* and *identity_hate* classes properly.

With above results, it was not prudent for us to pursue T5 due to no improvement in the metrics.

4.5 XLNet

Once T5 didn't give an improvement over Bert, we examined models that use a different approach to pre-training and implemented XLNet (Yang et al., 2019), which is an auto-regressive based model that does a better job of learning dependency pairs.

We ran XLNet on both the unbalanced and balanced data, for 6 epochs on each.

We found that XLNet gave us a modest improvement over Bert, both with balanced and unbalanced data.

5 Results and Discussion

Metrics were captured across multiple models that were tested as shown in Table 2 and Table 3. Each model was run for multiple epochs to ensure optimum fitting. We captured macro AUC-ROC scores, macro F1 score and AUC-ROC scores for different labels across different runs.

The baseline BERT model (with no re training) performed very poorly on the minority classes on both the unbalanced and balanced datasets. We fine-tuned the model by unfreezing multiple layers and trained BERT to learn the toxic data. As can be seen in the tables, there was a significant improvement in the F1, AUC-ROC scores. We also saw the minority classes performing well in this situation.

T5 was applied on unbalanced data and balanced data. The observations showed that T5 does worse on all of our metrics compared to BERT and XLNet with the exception of classifying *severe_toxic* data where it outperforms all the other models by a huge margin.

Similarly XLNet was applied on both unbalanced as well as balanced datasets. We observed that XLNet outperformed all of our attempts with other models for both the datasets.

Overall, the metrics across all the labels are more uniformly distributed when the models are trained on the balanced dataset. Referring to Figures 3 and 4 there is a clear indication of how each model shines given a balanced dataset.

Model	Macro F1	Macro AUC	Toxic AUC	Severe Toxic AUC	Obscene AUC	Threat AUC	Insult AUC	Id Hate AUC
BERT Baseline	0.20	0.56	0.66	0.50	0.63	0.50	0.61	0.50
BERT 2-layers	0.50	0.74	0.88	0.60	0.87	0.57	0.80	0.71
BERT 4-layers	0.55	0.78	0.90	0.59	0.86	0.71	0.88	0.72
BERT 6-layers	0.58	0.85	0.91	0.83	0.89	0.85	0.87	0.73
BERT 12-layers	0.49	0.77	0.90	0.73	0.89	0.49	0.88	0.73
T5	0.07	0.59	0.53	0.94	0.51	0.58	0.50	0.50
XLNet	0.60	0.82	0.90	0.69	0.87	0.75	0.87	0.85

Table 2: Metrics - All Models trained with original unbalanced data

Model	Macro F1	Macro AUC	Toxic AUC	Severe Toxic AUC	Obscene AUC	Threat AUC	Insult AUC	Id Hate AUC
Bert Baseline	0.05	0.51	0.57	0.5	0.50	0.49	0.50	0.50
Bert 2 layers	0.42	0.86	0.85	0.79	0.87	0.91	0.83	0.89
Bert 4 layers	0.44	0.87	0.60	0.88	0.87	0.92	0.83	0.88
Bert 6 layers	0.50	0.83	0.87	0.73	0.83	0.88	0.79	0.90
Bert 12 layers	0.51	0.89	0.89	0.87	0.88	0.90	0.87	0.91
T5	0.12	0.61	0.77	0.92	0.52	0.49	0.50	0.50
XLNet	0.42	0.90	0.89	0.83	0.90	0.95	0.89	0.93

Table 3: Metrics - All Models trained with augmented balanced data

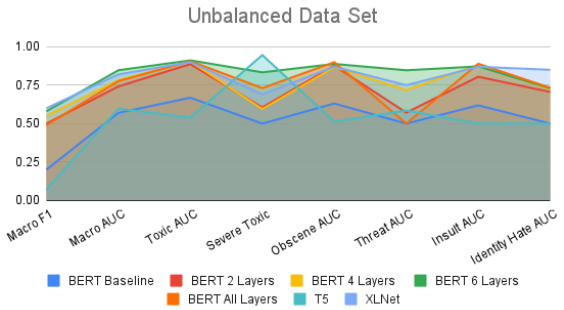

Figure 3: Unbalanced Data set

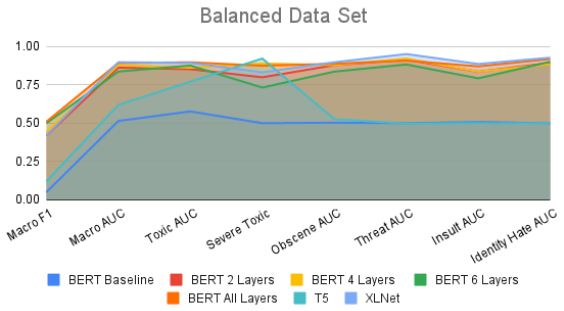

Figure 4: Balanced Data set

6 Conclusion

In conclusion we found that the benefits of a pre-trained large language model are limited by the vocabulary that it was trained on. With Bert being trained only on Wikipedia data it hasn't seen the same level of toxic content as in our training dataset which makes it harder to classify toxic comments correctly. T5 was pre-trained on a large corpus of data (C4) where it would have picked up more toxic content. The metrics for *severe_toxic* in Table 3 as well as 2 clearly shows that it is biased towards that label due to some pre-learning whereas few other classes were left out. With XLNet outperforming the other two models in the metrics, we can clearly say that XLNet is the best at multi-label classification on toxic data.

7 Future work

Our experiments with different models were based on the Jigsaw Toxic Comment dataset and some augmentation was done to balance it. Our source dataset was labeled by human annotators, who can bring their own biases. We have not focused on the mitigation of these biases and they may be reflected in our models. This is one area we can look at in the future.

For future work, we are inspired by Baldini et al. (2021) to evaluate and optimize our model for fairness across sensitive groups, like age, gender and race / ethnicity.

We have seen that each of the model has its own area of strength. So, we would like to create an ensemble model derived from the strengths shown by BERT, T5 and XLNet.

References

Adams, C., Sorenson, J., Elliott, J., Dixon, L., McDonald, M., and Cukierski, W. (2017). Kaggle toxic comment classification challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge.

Baldini, I., Wei, D., Ramamurthy, K. N., Yurochkin, M., and Singh, M. (2021). Your fairness may vary: Pretrained language model fairness in toxic text classification. *arXiv preprint arXiv:2108.01250*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dinakar, K., Reichart, R., and Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 11–17.

Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Haworth-Roberts, A. (2002). Bullying don't suffer in silence: an anti-bullying pack for schools. https://webarchive.nationalarchives.gov.uk/ukgwa/20050302035856/http://dfes.gov.uk/bullying/pack/02.pdf. [Online; accessed Apr-2023].

Jaunjale, P. (2020). T5 - Multi Label Classification. https://www.kaggle.com/code/prithvijaunjale/t5-multi-label-classification/notebook. [Online; accessed Apr-2023].

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Rastogi, C., Mofid, N., and Hsiao, F.-I. (2020). Can we achieve more with less? exploring data augmentation for toxic comment classification. *arXiv preprint arXiv:2007.00875*.

Singh, T. (2020). How much does pre-trained information help? partially re-initializing bert during fine-tuning to analyze the contribution of layers.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., and Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.