Personalized Censorship with Prototypical Networks

Abstract

This paper presents the development and optimization of personalized, user-adaptive text classification constructed using a hybrid approach which includes Bidirectional Encoder Representations from Transformers (BERT) for feature extraction and a prototypical network for an adaptive few-shot learning text classification model. Using this hybrid architecture, the aim is to classify potentially toxic or sensitive textual media that adapts to user preferences while maintaining accuracy and scalability. BERT powerfully embeds contextual meaning to convey potential nuance through the embeddings in which the prototypical-network will then classify. Using a balanced toxicity dataset of toxic (or sensitive) and nontoxic comments, the model has demonstrated a potential in enabling both generalized and personalized moderation to media. This research highlights the potential of combining pretrained language models and prototypical networks for quickly adaptable and efficient moderation, making it highly suited for real-world application.

1. Introduction

Social media has become an essential part of many individuals' lives. Concurrently, incidents of online hate speech and cyberbullying have exponentially risen in recent years [1] [2]. Therefore, this project seeks to develop a model fine-tuned to adapt to individual user preferences, censoring an appropriate level of potentially sensitive content.

The need to censor harmful online content has become increasingly prominent, especially as more at-risk populations—including children, and individuals with mental health conditions—are more vulnerable to its damaging effects. Exposure to this online hate speech, cyberbullying, and other harmful content has been shown to correlate with psychological outcomes such as increased anxiety and depression, especially among these vulnerable populations [3]. This is why developing censorship mechanisms to identity and limit exposure is important in reducing these hurtful effects. The censorship of such sensitive content poses significant challenges, as colloquialisms and informal language often express these thoughts, and what constitutes "sensitive content" (such as controversial, emotive, or hateful material) varies widely between individuals. This variability complicates the application of generic censorship algorithms, which may become overly restrictive. Currently, most modern solutions require extensive training and cannot adapt quickly, Therefore, a network is needed that is able to adapt easily without extensive retraining being needed.

There are many potential application areas for both individual users and businesses. For the day-to-day user, this model could be used as a browser extension, enabling personalized content censorship to fit their needs. For businesses, this functionality could be integrated as a platform feature, allowing companies to offer their users adaptive content censorship tools which fit to individual user preferences.

2. Related Work/Lit Survey/Background

Many studies have been conducted in toxic comment classification, including those utilizing BERT [4] and prototypical networks [5]. However, these methods were utilized differently. Nuthalapati et al. [6] conducted a study in which BERT was fine-tuned on the Jigsaw Toxic Comment Classification dataset [7] and achieved 94% precision at identifying toxic comments. However, this was a more generic algorithm which was not fine-tuned to a user preference and is prone to the issues mentioned earlier, however this method represents the underlying feature extraction that will be used. Nghiem et al. [8] found that using prototypical networks, while only utilizing 10% of the same dataset, a precision score of 75% could be achieved when classifying toxic comments. Both of these approaches communicate that toxic comment classification is practical using these methods.

3. Approach

To approach this, a hybrid method will be proposed that combines the pre-trained BERT embeddings with a prototypical network. To start, we will fine-tune BERT to better represent toxicity or sensitive content within our input.

3.1. Data Preparation

The data is preprocessed to include a balanced amount of toxic and non-toxic comments. Optionally, there is an option to include gathering comments which meet a specific active toxicity level (how many categories the comment is toxic in). This allows evaluation of both general and specific toxicity classification.

3.2. Embedding Extraction & Fine-tuning BERT

The input comments are processed through a pre-trained BERT model, which will create embeddings to serve as input features for the prototypical network. This is done to capture the nuanced meanings of toxic behavior as previously stated. As we are specifically capturing toxic behavior, the BERT model may be fine-tuned. The "distilbert-base-uncased" model was used, both standard and fine-tuned. To fine-tune these models, the model will be evaluated on how well it could preform bi-classification on our data, given the two classes being toxic and non-toxic.

3.3. Few-Shot Prototypical Learning

The episodic training method described in the introduction paper is employed, where each episode performs a `k_shot` and `q_query` task to optimize the model for minimal-data scenarios. Prototypes for each class (toxic and non-toxic) are computed by taking the mean of the embedded support set. Classification is then performed by comparing the query embeddings to these prototypes using a distance metric. L2 distance is used as it is shown to outperform other distance metrics such as cosine similarity.

3.4. Architecture

The model is architected as a fully connected layer, followed by a ReLU activation layer, along with a 20% dropout regularization layer, and finally another fully connected layer with layer normalization.

4. Experiments

The data that was used is the Jigsaw Toxic Comment Classification dataset. This was used as it has a large number of hand-labeled Wikipedia comments that have been posted. These comments were classified for the following categories: toxic, severe_toxic, obscene, threat, insult, identity_hate. An active toxicity level would be labeled with a one, and an inactive toxicity label would be a zero. If there are no active toxicity levels, then it would be labeled as non_toxic.

4.1. BERT

To fine-tune BERT, specifically the "distilbert-base-uncased" model, we experimented with several learning rates. The learning rates that were considered are $\lambda = 0.001$ (see Figure 1), 0.00005 (see Figure 2), and 0.000008 (see Figure 3).

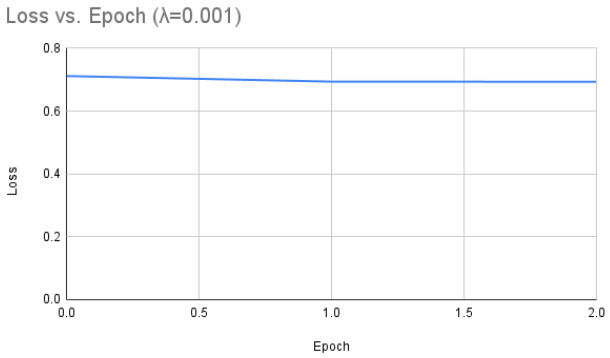

Figure 1. Learning rate of 0.001, final loss ≈ 0.6932.

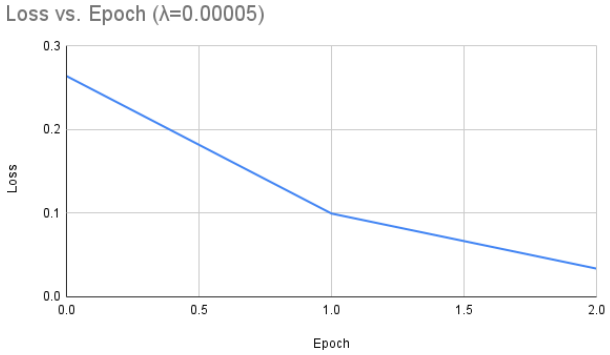

Figure 2. Learning rate of 0.00005, final loss ≈ 0.03376.

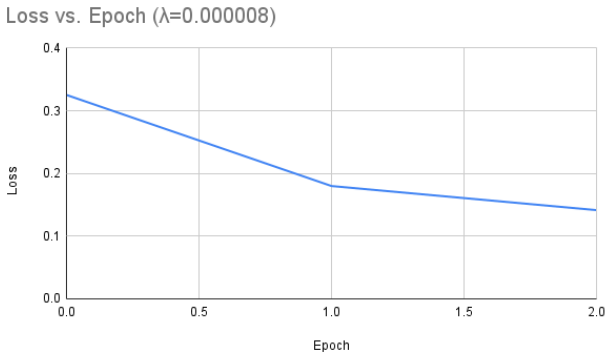

Figure 3. Learning rate of 0.000008, final loss ≈ 0.1415.

Here, we can see that figure 2 has the lowest loss.

4.2. Prototypical Network Training

The first experiment that was performed on the prototypical model was to tune the learning rate of the model. The learning rates that were considered are $\lambda = 0.001$ (see Figure 4), 0.0001 (see Figure 5), and 0.00001 (see Figure 6). The other parameters for this experiment being $k\_shot = 4$, $q\_queries = 9, epochs = 25$ along with using the standard "distilbert-base-uncased" BERT model. We can then plot the loss vs the current epoch as shown in the figures below.

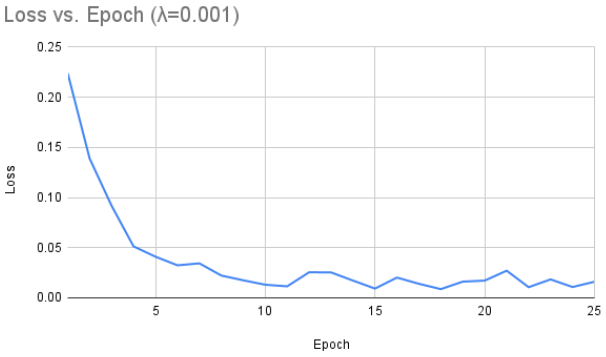

Figure 4. Learning rate of 0.001, final loss ≈ 0.0160373.

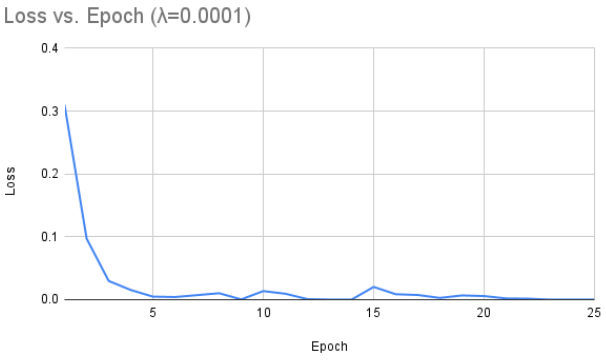

Figure 5. Learning rate of 0.0001, final loss ≈ 0.000021.

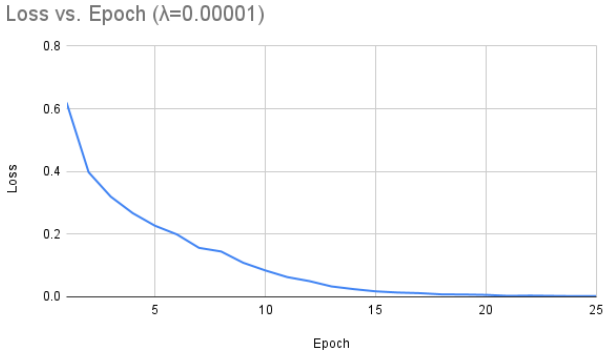

Figure 6. Learning rate of 0.00001, final loss ≈ 0.002286.

We can see that figure 5 has the lowest loss.

4.3. Prototypical Network Evaluation

The prototypical model was evaluated on balanced test sets. Accuracy was measured across the normal set of test data and along with data classified by the active toxicity column. Four experiments were performed to identify which would yield the highest accuracy. All experiments use the parameters `k_shot` $= 4$, `q_queries` $= 9$, and `epochs` $=$ 25 with the BERT model being "distilbert-base-uncased". Along with these parameters, the best learning rate will be used that was identified earlier. To enable better generalization so that the model will not be oversensitive to what many would not consider toxic or sensitive, the model can be trained on data that meets a minimum active toxicity level. Both of the experiments used with this method were performed at a minimum toxicity level of three.

Note: There is not enough data for the severe_toxic and threat columns as they are essentially always paired with one of the other toxicity categories, creating not enough data to evaluate only these toxicity labels.

4.3.1 Non-Fine-tuned BERT

This experiment utilizes a default, non-fine-tuned model of BERT with data of any active toxicity level.

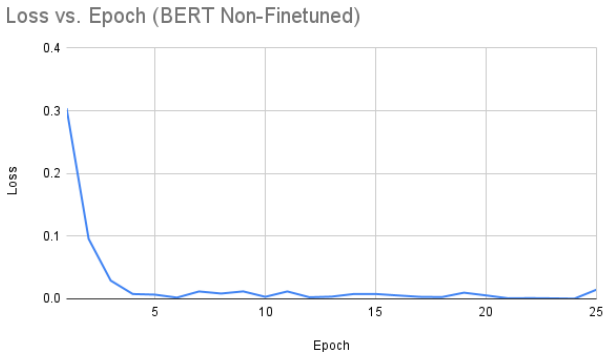

Figure 7. final loss ≈ 0.01459.

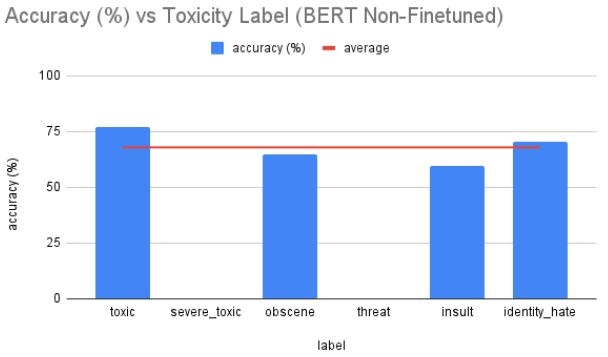

Figure 8. average accuracy of $\approx 67.905\%$.

4.3.2 Fine-tuned BERT

This experiment utilizes the best fine-tuned model of BERT that was previously identified and data of any active toxicity level.

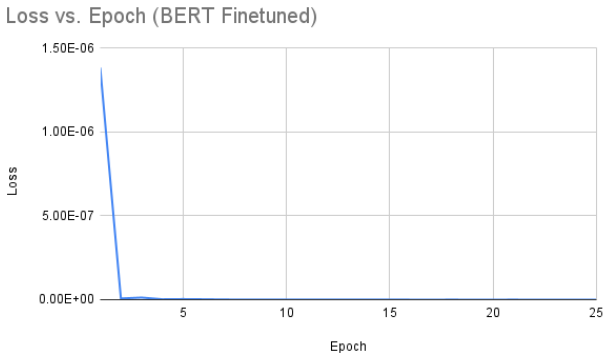

Figure 9. final loss $\approx 1.66 \times 10^{-11}$.

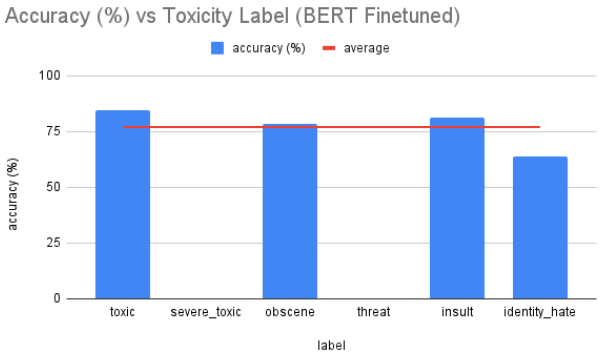

Figure 10. average accuracy of $\approx 77.03\%$.

4.3.3 Non-Fine-tuned BERT w/ Active Toxicity Level

This experiment utilizes a default, non-fine-tuned model of BERT with data that is balanced between non_toxic and an active toxicity level of 3.

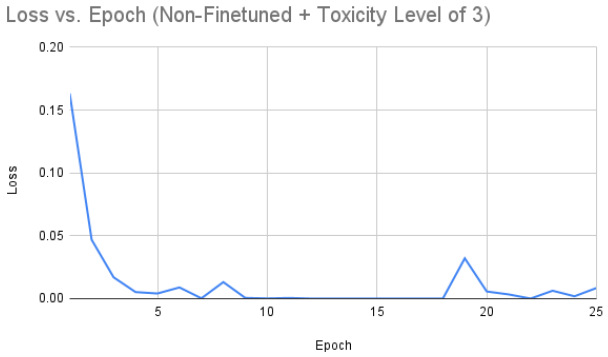

Figure 11. final loss ≈ 0.0084.

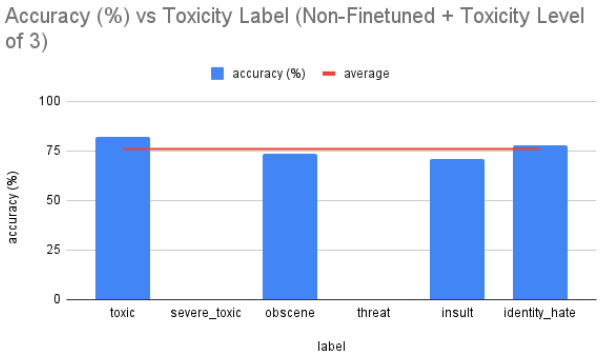

Figure 12. average accuracy of $\approx 75.99\%$.

4.3.4 Fine-tuned BERT w/ Active Toxicity Level

This experiment utilizes the best fine-tuned model of BERT that was previously identified and data that is balanced between non_toxic with an active toxicity level of 3.

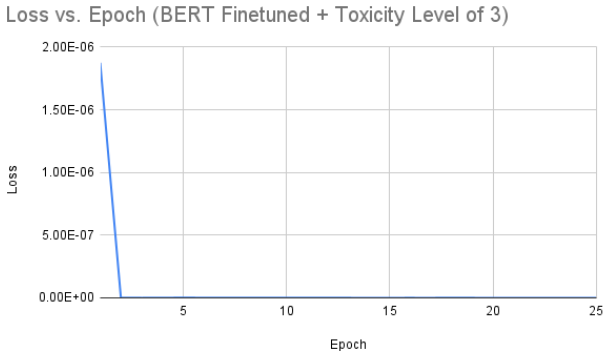

Figure 13. final loss ≈ 0.

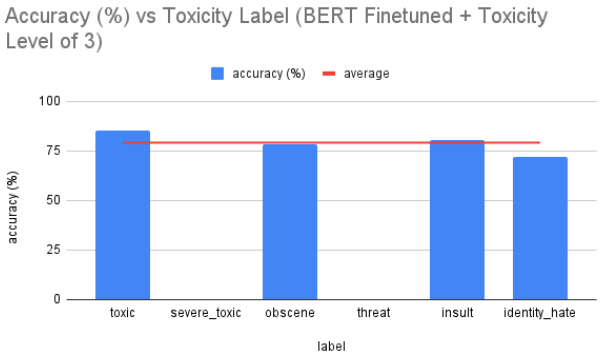

Figure 14. average accuracy of $\approx 79.185\%$.

These figures indicate that the prototypical network learns quickly and efficiently. Along with this, it is shown that having both a fine-tuned BERT model along with an active toxicity level of three yields the highest average accuracy.

4.3.5 Practical Use Testing

To attempt and evaluate how well this performs with queries not from the dataset and in real-world applications, we will create our own data to test. With an empty support query, we can first run our control cases. We will see what the model predicts "hello, how are you?", "circles are bad" and "squares are good" to be. The model outputs a 0 for all three, or that none of them are sensitive / toxic. With a support query of: "circles are bad" and "squares are good" we can start some very low-data testing. For now, we'll label the first support query as toxic (or a 1) and the second as nontoxic. We can input "hello, how are you?" as our query and run the model to evaluate. The model outputs that it's not sensitive / toxic. However, when we run our query "circles are bad" it now returns that it is toxic and or sensitive. The "squares are good" query prediction hasn't changed. This suggests that the model can correctly classify queries that are identical to the support queries. Lets now try some similar queries, these being "circles are terrible", "the worst shape is a circle", "squares are terrible" and "the worst shape is a square". Testing these with our model shows that it predicts sensitive for all of them, which is correct. We can also swap the labels around, so that "circles are bad" is not offensive, and "squares are good" is. Once again, queries that are identical to the support queries give the correct prediction. However, for our set of similar queries, we get: nontoxic, nontoxic, toxic and toxic. This is likely ideal, as the latter two queries are stating that something is bad / terrible when our support query doesn't explicitly say that is not offensive.

However, further testing shows that words that are often related to controversial issues more than other words often make a query predicted as sensitive, even if the entire query often wouldn't be considered sensitive. This is not ideal and might indicate a possible bias in the data which was not accounted for.

5. Conclusions

With nearly 80% accuracy, it is possible to predict whether a piece of text is either toxic or nontoxic while adapting to meet user needs using a hybrid of a fine-tuned BERT model and a prototypical network. Future ideas for the BERT feature extractor would be trying out different models of BERT along with fine-tuning these models more. For the prototypical model, some ideas would be trying different `k_shot` and `q_queries` values along with some stronger regularization or possibly more training data (400 data points were used) to prevent the overfitting that is seen in some of the plots (mainly figure 13 with 0 loss but only $\approx 80\%$ accuracy shown by figure 14). I have realized that there may often be unseen bias within the data that can affect how the model performs.

References

[1] Anti-Defamation League. Online hate and harassment 2023. Anti-Defamation League, June 2023. Accessed: 2024-09-24. 1

[2] Pew Research Center. Teens and cyberbullying 2022. Pew Research Center, December 2022. Accessed: 2024-09-24. 1

[3] Aiman El Asam and Adrienne Katz. Vulnerable young peo-

ple and their experience of online risks. *Human–Computer Interaction*, 33(4):281–304, 2018. 1

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019. Accessed: 2024-09-25. 1

[5] J. Snell, K. Swersky, and R. S. Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017. Accessed: 2024-09-26. 1

[6] Parthav Nuthalapati, Srinivas Abbaraju, G. Varma, and Sitanath Biswas. Cyberbullying detection: A comparative study of classification algorithms. *International Journal of Computer Science and Mobile Computing*, 13:1–12, 02 2024. 1

[7] C. J. Adams, J. Sorensen, J. Elliott, L. Dixon, M. McDonald, Nithum, and W. Cukierski. Toxic comment classification challenge, 2017. Accessed: 2024-09-30. 1

[8] Huy Nghiem, Umang Gupta, and Fred Morstatter. "define your terms" : Enhancing efficient offensive speech classification with definition, 2024. 1