ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts

Thanh Thi Nguyen[1], Campbell Wilson[1] and Janis Dalins[2]

[1]AiLECS Lab, Faculty of Information Technology,
Monash University, Melbourne, Australia

[2]AiLECS Lab, Australian Federal Police, Melbourne, Australia

Abstract.
This paper proposes an approach to detection of online harmful content using the open-source pretrained Llama 2 model, recently released by Meta GenAI. We fine-tune the LLM using datasets with different sizes, imbalance degrees, and languages. Based on the power of LLMs, our approach is generic and automated without a manual search for a synergy between feature extraction and classifier design steps like conventional methods. Experimental results show a strong performance of the proposed approach, which is proficient and consistent across three distinct datasets with five sets of experiments. This study's outcomes indicate that the proposed method can be implemented in real-world applications (even with non-English languages) for flagging sexual predators, offensive or toxic content, and hate speech in online discussions and comments to maintain respectful digital communities.

1 Introduction

Researchers have attempted to develop algorithms that are able to identify predatory behaviours and offensive language in online conversations. Those approaches however do not normally constitute a one-size-fits-all model. This implies that a suggested model could exhibit strong performance on one dataset while potentially demonstrating poor performance on a different dataset. For example, a method may execute competently on English-language text data, but perform imperfectly on non-English data, or a method may carry out effectively on a balanced dataset, but perform deficiently on an imbalanced dataset. Likewise, a model may be appropriate for short-text data analysis but may not be suitable for processing and analysing long-text data.

In addition, existing text classification approaches generally must involve a combination of two main steps, namely feature extraction and classifier design. Researchers may need to traverse through an extensive list of combinations of these two steps to find out the best synchronization or synergy between them. For example, experiments in [1] were conducted with 16 combinations between four feature extraction methods and four classifiers. Similarly, the work in [2] had to experiment with 17 classifiers.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

2 Related Works and Our Contributions

Given the extensive development of text classification methods, we focus on related works within the field of detecting *online harmful content* such as sexual predatory conversations, grooming, offensive, cyberbullying texts or hate speech based on *recent language models*. Language models such as transformers can be used to extract features or embeddings, which are then fed into a classifier. They can also be applied to text classification directly by adding a sequence classification head (e.g., a linear layer) on top of them. The former approach is exhibited in [1] where BERT and RoBERTa are used in feature/embedding extraction methods. The latter approach is demonstrated in [3].

In contrast to previous works, this paper introduces an approach to text classification by fine-tuning a modern open-source *large language model*, i.e., a pretrained Llama 2 model, for detecting online sexual predatory conversations and abusive language. The contributions of this paper are fourfold: 1) we fine-tune a Llama 2 model as a one-size-fits-all approach, i.e., our method works across different languages, data sizes, imbalance degrees, and does not require the use of text pre-processing techniques that are widely used in other approaches such as lemmatizing, stemming, emojis transcribing, lower-casing, and spell correcting; 2) we provide a review of literature related to detection of online harmful content using language models; 3) we perform a series of experiments on different datasets and compare our approach against state-of-the-art methods with extensive discussions; 4) we highlight the advantages of Llama 2 fine-tuning with both English and non-English languages even though Llama 2 models were pretrained mostly on English (Western demographic) data.

3 Llama 2-based Text Classification Approach

Llama 2 architecture and pretraining process involve the use of RMSNorm pre-normalization, SwiGLU activation function, and rotary positional embeddings. The LLaMA tokenizer uses the bytepair encoding based on the implementation from SentencePiece with a vocabulary size of 32k tokens [4].

Our proposed fine-tuning approach is depicted in Fig. 1. The original weights of the 7B-parameter Llama 2 model are converted to the Hugging Face transformer format to take advantage of the Hugging Face fine-tuning tools. The converted model is encapsulated within the PEFT-LoRA framework, which implements the Low-Rank Adaptation (LoRA) method [5] based on the Parameter-Efficient Fine-Tuning (PEFT) library. The LoRA approach was inspired by the structure-aware intrinsic dimension method [6] and it introduces a different way to execute low-rank fine-tuning. In LoRA, the parameter update for a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ is specified by a product of two low-rank matrices W_A and W_B:

$$\delta W = W_A W_B \tag{1}$$

where $W_A \in \mathbb{R}^{d\times r}$ and $W_B \in \mathbb{R}^{r\times k}$ are matrices of trainable parameters, and the rank $r \ll \min(d, k)$ [7]. The pretrained model parameters W_0 are frozen during

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

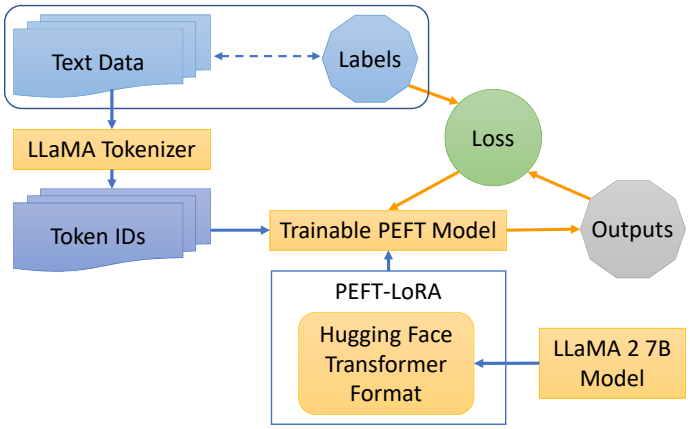

Fig. 1: The approach to fine-tuning the pretrained Llama 2 model for text classification. Text data are fed into the LLaMA tokenizer and converted to token IDs lists, which are then used as inputs to the trainable PEFT model. This model is created by converting the original LLaMA weights to the Hugging Face transformer model, which is then loaded through the PEFT-LoRA framework.

training and do not receive gradient updates. Both W_0 and $\delta W = W_A W_B$ are multiplied with the same input [5]; hence, the forward pass for $h = W_0 x$ is altered as:

$$h = W_0 x + \delta W x = W_0 x + W_A W_B x \tag{2}$$

After training, the trainable parameters can be combined with the original weight matrix W_0 by adding the matrix $W_A W_B$ to W_0. This approach allows the trainable matrices W_A and W_B to be trained to adapt to the new data while reducing the overall number of updates. As gradients for original pretrained weights are not computed, the GPU memory requirement is reduced significantly, leading to a fast and efficient LoRA-based fine-tuning. Our fine-tuning approach uses the cross-entropy loss function between network's output logits $\hat{y}$ and target y, which are labeled as "Outputs" and "Labels" in Fig. 1.

4 Datasets and Performance Metrics

We aim to evaluate and validate the proposed Llama 2 approach using datasets with different languages and properties such as data size, imbalance degree, and text length. Experiments in this study are therefore performed using three datasets: the PAN12 dataset [8] for detecting sexual predatory chats, and the Roman Urdu and Urdu datasets [9, 2] for abusive language detection.

The PAN12 competition dataset contains hundreds of thousands of conversations, described in detail in [8]. Each whole conversation was labelled as predatory or non-predatory.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

More specifically, the training set after preprocessing includes 952 predatory and 8,522 non-predatory conversations, while the testing set contains 1,698 predatory and 20,024 non-predatory conversations.

We acknowledge the ethical concerns around the use of data obtained through deceptive impersonation such as those from Perverted Justice, and do not in any way condone this methodology. However, we have used this dataset to facilitate comparison of our method with other work in this area.

The Roman Urdu dataset contains 147,180 comments on multiple YouTube videos and publicly available at: https://github.com/shaheerakr/roman-urdu-abusive-comment-detector. The comments were in Roman Urdu and labelled as abusive or non-abusive. Similar to the previous works in [9, 2], ten thousand data samples consisting of 5,000 abusive and 5,000 non-abusive comments were randomly extracted from this dataset to evaluate our proposed method.

The Urdu dataset was created by manually collecting YouTube comments in Urdu. This dataset is available at: https://github.com/pervezbcs/Urdu-Abusive-Dataset with 2,170 comments, which were annotated by local speakers as either abusive or non-abusive.

The following metrics are used to measure performance of competing methods: accuracy, true positive rate (TPR), false positive rate (FPR), and F_β with β being equal to 0.5 (i.e., $F_{0.5}$ score) and 1.0 (i.e., F_1 score or F-measure). The F_1 score gives equal weights to Precision and Recall while the $F_{0.5}$ score emphasizes Precision, putting more attention on minimizing false positives rather than minimizing false negatives.

5 Results and Discussions

With the aim towards a one-size-fits-all method, we set the same value across all experiments for each of the parameters in this study. The maximum sequence length is set at 128 tokens. The LoRA attention dimension (the rank of the update matrices) is equal to 8, while the LoRA alpha (the alpha parameter for LoRA scaling) is set to 16.

5.1 Sexual predatory chat detection using PAN12 dataset

A comparative summary of the performance metrics for different methods used in detecting online sexual predatory chats is displayed in Table 1. The most recent work on this topic, i.e., Borj et al. [1], achieved a high accuracy of 0.99, an F_1 score of 0.97, and an $F_{0.5}$ score of 0.98 (i.e., values in the row "Borj et al. [1] - Fusion" in Table 1). Our Llama 2 method stands out with excellent accuracy with a value of 1.00, and strong F_1 and $F_{0.5}$ scores both of 0.98.

The best result in [1] (row "Borj et al. [1] - Fusion") was obtained by trialling three fusion approaches at both score-level fusion and decision-level fusion on outcomes of 16 individual models. In contrast, our approach does not have to exhaust various models or all possible combinations of those models.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Table 1: Detecting sexual predatory chats using the PAN12 dataset

Competing Methods	**Performance Metrics**		
	Accuracy	F_1	$F_{0.5}$
Villatoro-Tello et al. (shown in [1])	0.92	0.87	0.93
Fauzi & Bours (shown in [1])	0.95	0.90	0.93
Borj & Bours (shown in [1])	0.98	0.86	-
Bours and Kulsrud (shown in [1])	-	0.94	0.97
Borj et al. (shown in [1])	0.99	0.96	-
Ebrahimi et al. (shown in [1])	0.99	0.77	-
Borj et al. (shown in [1])	0.99	**0.99**	0.94
Borj et al. [1] - Individual	0.99	0.96	0.96
Borj et al. [1] - Fusion	0.99	0.97	0.98
Our Llama 2 Approach	**1.00**	**0.98**	**0.98**

5.2 Abusive texts detection using Roman Urdu and Urdu datasets

For each of the two abusive text datasets (i.e., Roman Urdu and Urdu), we carried out two sets of experiments for comparisons with existing works: one for the training-testing data split ratio of 80-20, and another for the ratio of 90-10.

Table 2: Detecting abusive texts using the Roman Urdu and Urdu datasets

Training-Testing Data Ratio (%): 80-20

Datasets	**Roman Urdu**				**Urdu**			
Methods	***Acc.***	F_1	***TPR***	***FPR***	***Acc.***	F_1	***TPR***	***FPR***
2-Layer LSTM [9]	-	85.7	85.7	14.3	-	91.1	91.1	8.9
2-Layer BLSTM [9]	-	86.2	85.8	14.2	-	92.1	92.1	7.9
1-Layer LSTM [9]	-	88.2	87.6	12.4	-	93.5	93.5	6.5
CLSTM [9]	-	88.6	88.1	11.9	-	94.3	94.2	5.8
CNN [9]	-	**91.6**	91.4	8.6	-	**96.2**	96.1	**3.9**
Llama 2 Approach	**99.3**	**99.3**	**98.9**	**0.4**	**95.9**	**95.9**	**96.4**	4.7

Training-Testing Data Ratio (%): 90-10

Datasets	**Roman Urdu**				**Urdu**			
Methods	***Acc.***	F_1	***TPR***	***FPR***	***Acc.***	F_1	***TPR***	***FPR***
SVM-Polynomial [2]	97.7	97.7	-	-	95.5	95.5	-	-
Rule-JRip [2]	98.2	98.2	-	-	92.8	92.8	-	-
SimpleLogistic [2]	98.3	98.3	-	-	**95.9**	**95.9**	-	-
REPTree [2]	98.9	98.9	-	-	94.9	94.9	-	-
LogitBoost [2]	**99.2**	**99.2**	-	-	94.9	94.9	-	-
Llama 2 Approach	**99.5**	**99.5**	**99.2**	**0.2**	**96.3**	**96.6**	**97.4**	5.0

Across all four experiments as presented in Table 2, our method demonstrates an excellent performance compared with state-of-the-art methods in [2, 9]. For example, our method performs best across the two datasets in the experiments using the 90-10 training-testing data split ratio, i.e., the bottom part of Table 2. On the other hand, the best method in [2] on the Roman Urdu dataset is the LogitBoost model. That method however is not the best method on the Urdu dataset in [2] (which is the SimpleLogistic model). Furthermore, the CNN model in [9] demonstrates good performance on the Urdu dataset but performs poorly on the Roman Urdu dataset compared with most models in [2]. This highlights the importance of our LLM-based approach as it demonstrates stable and consistent performance across different characteristics of data and languages.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

6 Conclusion and Further Work

In this paper, we have fine-tuned the pretrained Llama 2 LLM using the LoRA method for text classification with an interesting application in detecting online sexual predatory chat logs and abusive texts. Our proposed method is deemed a one-size-fits-all approach as it consistently delivered impressive results across various factors such as different languages, degrees of data imbalance, data sizes, and text lengths. We experimented our method on English, Roman Urdu and Urdu datasets and it demonstrated excellent performance in all tested languages. Even with a small number of non-English data samples used for fine-tuning, i.e., the Urdu dataset with only 1,736 training data samples, the LLM-based approach outperformed most of the traditional text classification methods.

Acknowledgement

This research/project was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

References

[1] Parisa Rezaee Borj, Kiran Raja, and Patrick Bours. Detecting online grooming by simple contrastive chat embeddings. In *9th ACM Int. Workshop on Security and Privacy Analytics*, pages 57–65, 2023.

[2] Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed Abdelmajeed, and Muhammad Tariq Sadiq. Automatic detection of offensive language for Urdu and Roman Urdu. *IEEE Access*, 8:91213–91226, 2020.

[3] Kanishk Verma, Maja Popović, Alexandros Poulis, Yelena Cherkasova, Angela Mazzone, Tijana Milosevic, Brian Davis, et al. Leveraging machine translation for cross-lingual fine-grained cyberbullying classification amongst pre-adolescents. *Natural Language Engineering*, pages 1–23, 2022.

[4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[6] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *59th Annual Meeting of the ACL*, volume 1, pages 7319–7328, 2021.

[7] Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.

[8] Giacomo Inches and Fabio Crestani. Overview of the international sexual predator identification competition at PAN-2012. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 30, 2012.

[9] Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed AbdelMajeed, and Tehseen Zia. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems*, pages 1–16, 2021.