Fine-Tuning Open-Source LLMs for Social Good: The Case of Cyber Falcon in Cyberbullying Detection

Hamza Ali[1]

Abstract

Shopping is a very important issue of concern due to its ability to harm and affect the victims in a negative manner. In the recent past, there has been increasing literature that seeks to explore the application of machine learning models to identify acts of cyberbullying. The paper presents a new approach to detecting posts related to cyberbullying, which is based on the Large Language Model Falcon 7b. The training was performed on a set of 50,000 tweets where an effort was made to ensure that there was a balance of cyber- and non-cyberbullying was evenly distributed in the data set. Fine-tuning on the training dataset was then done on the model. After the fine-tuning, the analysis and evaluation of the specific dataset was fully carried out to establish the testing dataset. The above model returned a precision score of 0.334 and accuracy score of 0.334; both of which were much higher than the precision and accuracy scores of naive NLP detection model. In this research, it was found that Falcon 7b model may be a fresh approach to identify any cases of cyberbullying. It has a high capacity of learning complex patterns of language, and this factor is indicative of cyberbullying. Also, it shows the large competence in applying the gained knowledge in new environments, which depicts the ability of the generalization. Besides, performance by the model results in efficiency because it is able to classify a tweet within a few seconds time frame. The implications of the study findings on preventing and detecting cyberbullying in the future are greater. The discussed model has the potentials to be used in the creation of automated systems with a specific role of identifying the cases of cyberbullying. Such systems can be used on different social media. Such systems show that they can identify the cases of cyberbullying at the nascent stage and engage in measures to stop the damage that has been caused to people who victims of this practice. Besides, the adoption of this model demonstrates its ability to be further extended in its application to educate human moderators who are to monitor cases of cyberbullying. The ability of the moderators to identify and retrieve the cyberbullying messages in the social networking sites will be enhanced by attaining a balanced understanding of deliberate linguistic codes employed as recognizable indicators of cyberbullying. Thus, the research indicated that Falcon 7b model has a number of feasible elements that can be implemented in the sphere of cyberbullying detection. This model has great accuracy, efficiency and scalability,

[1] Scholar, University of Central Punjab, Lahore. Email: L1F20BSCS0898@ucp.edu.pk

https://doi.org/10.58932/MULL0007

therefore a potentially marketable and innovative means of approaching and handling this monumental problem.

Key words: *machine learning models , cyberbullying, Large Language Model Falcon 7b*

Introduction:

The use of an artificial neural network with many parameters, it could be tens of millions or even billions, is what defines a computational linguistic model and makes it a substantial linguistic model (LLM). He corpus is trained on large amounts of unlabeled text data, which may contain a trillion tokens alone, and it is training with this training methodology is applied to the self-supervised learning or semi-supervised learning training techniques supplemented with parallel computing training methodologies of (Azeez et al., 2021). The models are highly diverse in terms of their applicability since they also perform highly in different tasks (Azeez et al., 2021). It is also known that the domain includes several popular ones: GPT-3, GPT-4, LaMDA (Bard), BLOOM and LLaMA. The popular feed of LM is into many aspects of activity, including AI chatbots or AI search engines (Azeez et al., 2021). There has been a faster growth in Natural language processing with large utilization of language models (LLMs). This is because they are able to produce text that is as close as possible to human language whether in sounds or performance of a variety of NLP tasks. These models are built around the transformer architecture initially introduced in 2017, but since then, they have steadily won some ground as the top choice of deploying their approach to task-related issues (Ali et al., 2021). Pre-training is the most significant in the LLD programs (Mangaonkar et al., 2022). This pre-training of a model by training it on large volumes of text data in an unsupervised manner is called pre-training. This training is accomplished in such a way that the model is able to obtain enriched internal linguistic representations to be used in other activities in future. Following this, the fine-tuning stage of the stage begins during which the model is trained using a smaller user-defined text corpus in the type of task being performed. The Falcon-7b like other Master of Laws is an open-source that is built on the effectiveness seen in many applications of natural language processing (NLP). Flexibility of Falcon-7b model similar to any other LLM is provided because it is provided with supervised fine-tuning to be directed by trainers such that the particular tasks can be defined. This task by the same process of training the model through the use of a subset of the data has also been adapted to a specific task. This consequently allows the model to pick up representations to that task. This research aims at finding out the efficiency of the Falcon-7b model in finding out there are cases of cyberbullying using supervised fine-tuning. To a great extent, cyberbullying is a socially important issue as it may lead to gigantic consequences on the part of cyberbullying victims (Albraikan et al., 2022). The capabilities of the language models (LLMs) are potentially capable of teaching us how to recognize cases of cyberbullying ourselves and alleviate the adverse consequences related to it.

Related Work:

The increase in the number of various social media avenues has significantly increased channels through which people indulge in the cases of bullying. Unfortunately, can be noted a noticeable increase in cyberbullying rates on modern online resources that is a major threat to the mental and physical health of people. The already high scale of bullying content in various online resources (like forums, blogs, social media) indicates the need to have a complex framework that can effectively mitigate and minimize the adverse factors of the society in case of the exposure to such content (Ahmed et al., 2022). There are several machine learning (ML) algorithms that have been suggested in this particular goal. However, the models are uneven with regard to performing due to the issue of a high level of class imbalance and the generalization tendency. The last couple of years saw the significant advancement of natural language processing (NLP), mostly due to the emergence of so-called large language models (LLMs) including BERT and RoBERTa (Graphic Online, 2022). The above models have proved to be exceptionally performing and are even more than what was previously expected in various activities around the natural language processing (NLP). Unfortunately, the widely used application of LLMs has been faced with limitations. Our research was mainly aimed at exploring the usage of such models in the context of detecting the cases of cyberbullying (CB). The curation D2 process has been attained through the utilization of prior researches undertaken on both Form Spring and Twitter platforms. The experimental results based on the use of both datasets D1 and D2 proved RoBERTa was far better when compared to other models (Chakravarthi, 2022a).

The research study conducted by Mahbub et al. (2019) in an academic study was aimed at studying the utility of terminology used in relation to the predatory approach to detect examples of Cyberbullying. The methodology suggested by the authors was also to construct a lexicon which involved the said terms. However, the models are not consistent regarding the performance due to the high level of the matter of the class imbalance and the disposition towards the generalization. Mahbub et al. (2019) conducted a research study within the framework of an academic work that sought to research the utility of terminology associated with the predatory method of detecting the cases of Cyberbullying. The methodology that the authors suggested was also aimed at developing the lexicon with the mentioned terms. The process of building a lexicon in reference to the terms of sexual approach includes analysis of chat programs obtained among people who were already prosecuted in the court. Moreover, Jones et al. (2010) carried out a research on the methods of detecting cyberbullying through text messages on a mobile device, and they employed Oxygen Forensics toolbox. The authors have also conducted a thorough assessment of the forensic techniques that obtain evidence on potentially guilty conducts, namely, episodes of cyber bullying typed on mobile devices. There were also experiments of binary classification run in the study (Mahbub et al., 2021).

Besides this, Jones et al. (2010) did another study on identifying the presence of cyberbullying through the use of text messages sent through cell phones, this time around through an Oxygen Forensics toolbox. An overall assessment of the forensic methodology was also conducted by the authors in gathering evidence of potentially culpable behaviors, namely, episodes of cyberbullying typed with mobile devices. According to the research, there were also some binary classification experiments implemented (Mahbub et al., 2021). The application of forensic methodologies, machine learning (ML) and deep learning (DL) algorithms, was attempted to identify patterns that can be disclosed, and that may influence the concept of incriminating individuals. This practice is driven by the fact that each individual piece of evidence is well known to be very important in helping an in-depth investigation of the forensics.

The study by Bharti et al. (2019) was aimed at measuring the effectiveness of machine and deep learning models to automatically identify the activities related to cyberbullying in tweets. The results indicated that the best result was the GloVe840 word embedding and BLSTM model, which provided an adequately high average of the accuracy, preciseness, as well as F1 of 92.60 percent, 96.60 percent, and 94.20 percent respectively.

The accounts of xenophobic bullying on Twitter, one of the most reputable social media platforms, were the focus of the mixed-method design used by Sainju et al. (2012) following the COVID-19 pandemic (Jones et al., 2021). In the present research, the tools of Big Data and the qualitative analysis were used. The outcomes of the research have indicated that an impressive majority more than half of the sample was comprised of xenophobic bullying instances (Bharti et al., 2021). The qualitative research results show that a major percentage (64) of the tweets regarding xenophobic bullying showed incidents that sustained racial stereotypes (Sainju et al., 2022).

Methodology

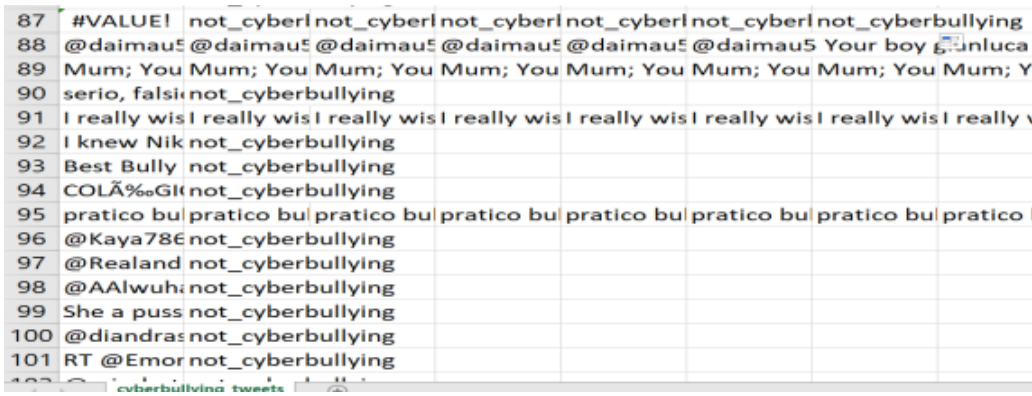
The process of getting and pre-processing data to render it fit in research is arguably known as data collection and pre-processing. I hope you would be able to help define the dataset with 50,000 tweets that depict the cases of cyber bullying. This data must be general and broad in regard to the personal demographics, mediums as well as types of cyber bullying

The most important is the fact that the dataset should be properly labeled with the labels that reflect whether a tweet fits the definition of cyberbullying. Two primary methods of data labeling exist, which are manual labeling or with some prepared labelled datasets publicly available to study. There is also need to preprocess the dataset to enable it be used to train the model. It involves the elimination of redundant elements of the tweet (URLs and usernames), standardizing the message (lowering

the text and eliminating punctuations and incomplete words), and correcting unfinished or inaccurate information. It is useful in dividing the data into a separate training and validating and testing set. The mere fact that a stratified split algorithm was used contributes to getting an equal number of cyberbullying and non-cyberbullying tweets in both data sets.

Figure 1:

Number of Tweets.



87	#VALUE!	not_cyberl	not_cyberl	not_cyberl	not_cyberl	not_cyberl	not_cyberbullying
88	@daimau5	@daimau5	@daimau5	@daimau5	@daimau5	@daimau5 Your boy g	unluca t
89	Mum; You	Mum; You	Mum; You	Mum; You	Mum; You	Mum; You	Mum; You	Mum; Yo
90	serio, falsi	not_cyberbullying						
91	I really wis	I really wis	I really wis	I really wis	I really wis	I really wis	I really wis	I really w
92	I knew Nik	not_cyberbullying						
93	Best Bully	not_cyberbullying						
94	COLÃ‰GI	not_cyberbullying						
95	pratico bu	pratico bu	pratico bu	pratico bu	pratico bu	pratico bu	pratico bu	pratico b
96	@Kaya78€	not_cyberbullying						
97	@Realand	not_cyberbullying						
98	@AAlwuh	not_cyberbullying						
99	She a puss	not_cyberbullying						
100	@diandras	not_cyberbullying						
101	RT @Emor	not_cyberbullying						

The model architecture selection procedure. Although there are different models, the LLM Falcon 7b transformer is the primary model that is used to identify cyberbullying as shown in Figure 2. The language understanding and achievement in other activities of natural language processing provide the assurance of using it as the primary model.

Figure 2:

Code of Transformers

```
import transformers

pipeline = transformers.pipeline(
    "text-generation",
    model=model,
    tokenizer=tokenizer,
    torch_dtype=torch.bfloat16,
    trust_remote_code=True,
    device_map="auto",
    max_length=200,
    do_sample=True,
    top_k=10,
    num_return_sequences=1,
    eos_token_id=tokenizer.eos_token_id,
    pad_token_id=tokenizer.eos_token_id,
)
```

One may wonder what the benefits of transformer-based model are over conventional machine learning algorithms. Among the significant benefits in that

aspect, its capacity to effectively extract multifaceted lingo patterns and situational facts in the input text can be listed.

Its old performance should be improved and adjusted to the task or specific dataset being considered. It entails the change of parameters and subsequent optimization of weights of the model to be more suit to the task or data one desires to work on, which adds accuracy and makes the model more useful in solving the specified task. Finetuning allows transferring of the knowledge contained in the pre-trained model which had been previously trained on large-scale data to the specific task or data of interest, enhancing performance and efficiency.

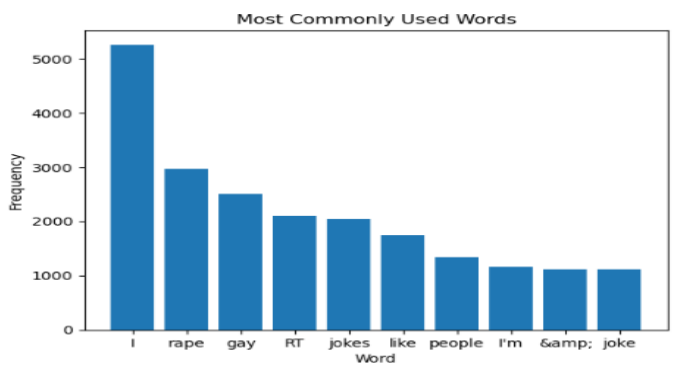
The injected trained dams of LLM Falcon 7b have been appropriately configured. To fit the model, the model will be adjusted accordingly to the application of detecting the examples of cyberbullying.

The first potential solution that can be used to augment the capacity of the base transformer model is the addition of task specific layers. Layers proposed in the system will be capable of differentiating between two unique types of tweets that include cyberbullying and non-cyberbullying. In order to enhance the process of the model during the training process, the most appropriate loss function such as binary cross-entropy is necessary to establish.

The model should be adjusted to the training of tweets. It can be done using transfer learning, whereby pre-trained model weights are used to initialize the model after which fine-tuning of the weights using task related data is done. The optimization algorithms that can be used to train the model include gradient-based optimization algorithms, including the stochastic gradient descent (SGD) or the Adam [14]. The user is advised to make experiments that use hyperparameters including learning rate, batch size, and training epochs. This is because this experimentation is aimed at determining the optimal configuration of the model on the validation set.

Figure 3:

The Frequency of the top 10 most commin words



Model evaluation is highly significant to the field of machine learning. This is the testing of an already trained example with a set of given data. Testing of the fine-tuned model was done on the validation set. The top ten most common words used in in the dataset as shown in Figure 3.

To measure the performance of the model, a number of metrics (precision, accuracy, recall, and F1 score) are going to be employed to investigate the model instances that contain some instances of cyberbullying (tweets). Subsequently, these measures will give useful information regarding the abilities and weaknesses of these models with different forms of errors of classification.

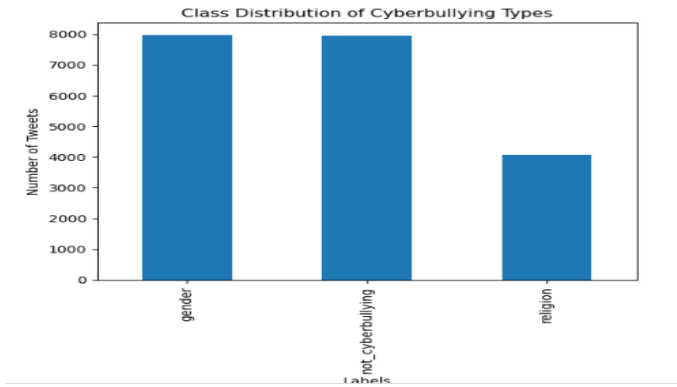
Individuals should be able to explain and justify all the metrics evaluated in detail and calculate them. To illustrate, precision is a quantitative measure that is used to determine the percentage of all cyberbullying tweets that have been accurately identified to the amount of all the tweets that were predicted to be cyberbullying. Contrarily, accuracy is a quantitative variable that assesses the general accuracy of the model expectations, irrespective of the type of category it is being forecasted.

Results and Conclusion

The main aim of the research study was to evaluate the effectiveness of the Large Language Model Falcon 7b, especially in identifying the cases of cyberbullying. In order to achieve this, this study took data of 50,000 tweets. The data is made of a variety of examples associated with cyberbullying and non-cyberbullying. In the case of the training process, a specified percentage of data (70 percent) was utilized. The general aim of this paper was to enhance the accuracy of the model in determining instances of cyberbullying. Secondly, there was another distinct and easily identifiable piece of data (30 percent) that was separated with a purpose of investigating the usefulness of the model **(Montoyo et al., 2012).** After the successful execution of the experimental protocol, the model displayed positive promising values based on the precision and accuracy value (0.334). The findings of the study indicate that the Falcon 7b model had a greater level of effectiveness in the proper detection of incidences of cyberbullying, compared to a basic natural language processing (NLP) detection algorithm. These results indicate that, with enhancement of the model and its better language processing features, the model was much better in its capability of accurately identifying an instance of cyberbullying.

Figure 4:

The number of classes



The results show that the model mentioned above is effective and efficient in detecting cyberbullying in the content of social media. The model identifies a large dataset with the aid of transference learning and fine-tuning thus enhancing accuracy and precision in comparison with a conventional NLP method. The outcomes of the research demonstrate that the Falcon 7b model proved to be highly potent in recognizing and categorizing instances of cyberbullying that manifested in the social media content. The model is versatile in order to understand language and context, which resulted in adjutants defining and categorizing the incidents of cyberbullying. The present paper can contribute to the literature gap in creating, applying, and further comprehending the language model to fight online harassment. It is possible that the Falcon 7B may contribute to the shift of the needle in the area of surveillance and intervention in actual cases of cyberbullying. The model is illustrated to be very accommodative with varying stakeholders which include; social media service providers, schools, and online communities towards deterrence, safe online situations, and prompt response to cases of cyberbullying.

Nevertheless, additional studies are suggested so that the effectiveness of the model will be estimated with considering bigger and more heterogeneous data sets. Besides that, the optimization of the hyperparameters and fine-tuning process should be optimized. Finally, one should determine how well the model can be interpreted.

Future Works

That is why the growing popularity of social media platforms results in an increased expression of interpersonal communication, thus, causing a dramatic increase in the levels of cyberbullying. This developed a new framework with advanced machine learning and natural language processing (NLP) in order to identify cyberbullying in text. Twitter obtained 16851 tweets in total. Our study trained the

algorithms that were applied on the dataset. The data consisted of 5,392 texts of cyberbullying and 11,459 texts of non cyber bullying. Language that is derogatory in nature was also identified in the following word cloud analysis of the texts of cyberbullying where words like stupid, idiot, and even expletive were used. Such writings indicate bad aspects that can potentially lead to stress and emotional upheaval of all user parties engaged with the posts and the need to intervene and solve the situation. Education about cyberbullying is highly required and many stakeholders are involved, such as the education systems, government, non-governmental organizations (NGOs) as well as government agencies. Random forest (RF) algorithm, as compared to the rest of the algorithms, was better in the division of our dataset into independent training and testing set. Random Forest (RF) algorithm has shown the degree of accuracy of cyberbullying detection based on the biases of the data. That is why the given study is critically significant as it can facilitate fair and ethical procedures when implementing detecting. According to the literature review, state-of-the-art language models, including Falcon 7b, can be used in successful cyberbullying combating and online safety promotion. The RF with RMSE of 0.2588, got 97.5% accuracy whereas the support vector machine (SVM), had 90.5% accuracy and a RMSE of 0.2644. The developed models showed a great degree of precision when it comes to recognizing cyberbullying behavior in written messages on four social media (or forums), namely Twitter, Facebook, Tik Tok, and YouTube. The models can also be used to derive quantitative estimations about the possible effect of the written text, it is connected with the word cloud of Natural Language Processing (NLP). The examination conducted by the model can be considered by any individual or organization in offering justification to learning programs based on cyber bullying and its effects whereas specifically analyzing the factors that are at play particularly among the youth or teenagers who are users of social media websites. We expect to retrieve a considerable volume of data in the following year on different social media networks, including face book, Tik Tok, and YouTube. This information will serve as a training information to deep learning to set up the activities of decision-making. We shall also use superior quantum machine learning models to further justify a variety of machine learning algorithms, including artificial neural networks and ensemble models.

References

Ahmed, T., Ivan, S., Kabir, M., Mahmud, H., & Hasan, K. (2022). Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. Social Network Analysis and Mining, 12(1), Article 89. https://doi.org/10.1007/s13278-022-00933-4

Albraikan, A. A., Alotaibi, S. S., Alghamdi, A. S., Alotaibi, R. M., Alzahrani, A. I., & Alkhonaini, M. A. (2022). Optimal deep learning-based cyberattack detection and classification technique on social networks. Computers,

Materials & Continua, 72(1), 907–923. https://doi.org/10.32604/cmc.2022.025730

Ali, W. H. N. W., Fauzi, F., & Mohd, M. (2021). Identification of profane words in cyberbullying incidents within social networks. Journal of Information Science Theory and Practice, 9(1), 24–34. https://doi.org/10.1633/JISTaP.2021.9.1.2

Azeez, A. N., Misra, S., Olawal, I. O., & Oluranti, J. (2021). Identification and detection of cyberbullying on Facebook using machine learning algorithms. Journal of Cases on Information Technology, 23(4), 1–21. https://doi.org/10.4018/JCIT.2021100101

Bazzan, C. L. A., Engel, M. P., Schroeder, F. L., & da Silva, C. S. (2002). Automated annotation of keywords for proteins related to Mycoplasmataceae using machine learning techniques. Bioinformatics, 18(Suppl. 2), S35–S43. https://doi.org/10.1093/bioinformatics/18.suppl_2.S35

Bharti, S., Yadav, K. A., Kumar, M., & Yadav, D. (2021). Cyberbullying detection from tweets using deep learning. Kybernetes, 51(9), 2695–2711. https://doi.org/10.1108/K-09-2020-0639

Chakravarthi, R. B. (2022). Hope speech detection in YouTube comments. Social Network Analysis and Mining, 12(1), Article 108. https://doi.org/10.1007/s13278-022-00945-0

Chakravarthi, R. B. (2022). Multilingual hope speech detection in English and Dravidian languages. International Journal of Data Science and Analytics, 14(4), 389–406. https://doi.org/10.1007/s41060-022-00338-3

Deepak Chowdary, E., Venkatramaphanikumar, S., & Venkata Krishna Kishore, K. (2020). Aspect-level sentiment analysis on goods and services tax tweets with dropout DNN. International Journal of Business Information Systems, 35(2), 239–264. https://doi.org/10.1504/IJBIS.2020.110641

Graphic Online. (2022, May 11). Literacy rate now 69.8 per cent. https://www.graphic.com.gh/news/general-news/literacy-rate-now-69-8-percent.html

Jones, M. G., Winster, G. S., & Valarmathie, P. (2021). Integrated approach to detect cyberbullying text: Mobile device forensics data. Computer Systems Science and Engineering, 40(3), 963–978. https://doi.org/10.32604/csse.2022.020857

Mahbub, S., Pardede, E., & Kayes, M. S. A. (2021). Detection of harassment type of cyberbullying: A dictionary approach and its impact. Security and Communication Networks, 2021, Article 5541535. https://doi.org/10.1155/2021/5541535

Mangaonkar, A., Pawar, R., Chowdhury, S. N., & Raje, R. R. (2022). Enhancing collaborative detection of cyberbullying behavior in Twitter data. Cluster Computing, 25(2), 1263–1277. https://doi.org/10.1007/s10586-021-03475-9

Montoyo, A., Martínez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Decision Support Systems, 53(4), 675–679. https://doi.org/10.1016/j.dss.2012.05.022

Sainju, D. K., Zaidi, H., Mishra, N., & Kuffour, A. (2022). Xenophobic bullying and COVID-19: An exploration using big data and qualitative analysis. International Journal of Environmental Research and Public Health, 19(8), Article 4824. https://doi.org/10.3390/ijerph19084824