Toxic or Not? A Deep Learning Approach to Comment Classification

Muhammad Hamza Amin
Faculty of Computer Science & Engg.
GIK Institute of Engg. Sciences & Tech.
Topi, Khyber Pakhtunkhwa, Pakistan.
u2022378@giki.edu.pk

***Abstract*—**Comment toxicity classification is an essential task for making the online environment healthier since toxic comments can provoke harassment among users and degeneration of the community. With time, the volume of comments to be classified only increases effectively to identify and address them.
Although numerous methods for comment classification exist, there seems to be a gap in terms of studying the effectiveness of deep learning, especially Long Short-Term Memory network and fine-tuned BERT models, regarding multi-class comment toxicity classification. The following paper aims to address the gap in identifying how these methods can help enhance classification accuracy. .

To conclude, two presented techniques: the TensorFlow and Keras-based LSTM model with good results and the fine-tuned BERT model with improved results for comment toxicity classification. Further, the findings show that the BERT model outperforms the LSTM model, which demonstrates the benefits of pre-trained language models. Therefore, the significance of this study is the way to efficient multiclass comment toxicity classification, which helps to provide a safe and respectful online environment on the platforms.

***Index Terms*—**fine-tuning, multiclass classification, LSTM, BERT, deep learning.

II. INTRODUCTION

Impactful information platforms online have made a breakthrough in our communication techniques. But at the same time, the frequent connection proves a substantial impediment, which is the spread of harmful and poisonous materials across the online society. These negativities come in various ways of insults, threats, hate speech, obscenity among others, hence forming an unfriendly and unproductive background.It is imperative to deal with toxic comments so as to enhance a healthy online conversation. Basically, they do not only prevent a meaningful conversation but also they can bring emotional distress and real-world harm. It is therefore necessary to come up with effective ways of identifying and mitigating comment toxicity for safer and inclusive online spaces.When it comes to this area, Kaggle's Toxic Comment Classification Challenge is quite a significant move. This is a good chance for people because there are already some large datasets with labeled comments where ?jal's of machine learning models that can be used in different fields have also being developed bomb. These models are very strong and so they will help you know many things which could not be detected before. Therefore, its importance cannot be underemphasized.

This report presents a detailed study of the problem, which involves examining the provided statistics as well as trying out different machine learning methods to ensure precise and effective assessment of comment toxicity.

A. Related Work

This studies report appears into this problem by means of reading the given records and investigating distinct system learning strategies with the intention to accomplish particular and efficient remark toxicity categorization. Considerable studies has targeted on mechanically detecting and classifying toxic feedback in Internet platforms. These strategies often employ supervised studying techniques, which use categorized statistics sets to train fashions the way to understand unique styles of harmful messages consisting of hate speech, insults, threats, or obscenities.
Some studies utilized conventional gadget mastering techniques which include Support Vector Machines (SVM) and Naive Bayes classifiers to yield advantageous algorithmic consequences for categorizing toxic remarks [1, 2]. In this context, deep learning models – most extensively, the Recurrent Neural Networks (RNNs) coupled with Long Short-Term Memory (LSTM) networks – have end up imperative attributable to their capacity for shooting sequence data inside textual access factors [3, 4]. Table I offers a summary document on diverse investigations accomplished heretofore where specific kinds of fashions.

B. Gap Analysis

Multiple shortcomings still exist in the domain despite significant headway in picking out and categorizing toxic comments. The present studies tend to focus mostly on binary classification– separating toxic from non-toxic comments. Nevertheless, online toxicity being very detailed needs a more intricate approach. It is important to realize individual kinds of toxicities like hate speech, threats or obscenities so that suitable mitigation strategies can be devised. At the same time, recent models often fail to cope with ever-changing nature of internet language which encompasses slang words, emerging phrases as well as culturally specific references among others things. On top of that, there is still no satisfactory way of dealing with shortcomings associated with multi-class classification models. This study seeks to fill these gaps by suggesting an original method which combines long short-term memory networks with fine-tuned BERT model for classifying toxic comments into six different categories.

C. Problem Statement

Following are the main research questions addressed in this study.

1. Can a multi-class classification model utilizing LSTM and fine-tuned BERT techniques effectively distinguish between six different types of toxic comments?
2. How do LSTM and fine-tuned BERT models compare in their performance on the multi-class classification of toxic comments?
3. How does the model handle characteristics of online language such as slang and cultural references that change over time?

D. Novelty of our work

In our approach to classifying toxic comments, we combine the strengths of the LSTM and BERT models. An LSTM network studies the sequential nature of language to understand the context and word relationships in each comment. Initially, the text is cleaned, sentences are tokenized and padded to ensure length consistency. Then, we add an embedding layer which maps words to vectors that represent their semantic relationships. An LSTM layer then processes this sequence of word vectors thereby capturing sequential information within comments. To reduce output dimensionality from LSTMs, global max pooling is carried out. Additional operations involve densely connected layers with ReLU activation.
This process is further enhanced through integration of fine-tuned BERT model providing profound contextual understanding. This gives insights on understanding text based on contexts, therefore helping models distinguish language nuances. A sigmoidal activation function at the last output layer makes binary predictions for the six categories of toxicity.

E. Our Solutions

In this report, we've got proposed an progressive method for remark toxicity classification that use LSTM and BERT models. We have incorporated an LSTM network to take gain of the language's underlying series and higher apprehend the remarks' context and the interactions between one of a kind phrases. A fine-tuned BERT version further enhances this and makes use of the pattern to pick out styles and make predictions primarily based on a extensive variety of statistics. As a end result, this approach of using LSTM and BERT models overperformed conventional techniques and lower back the best overall performance rate for the comment toxicity classification mission.

III. METHODOLOGY

A. Dataset

This work utilizes the publicly available dataset from the Kaggle 'Toxic Comment Classification Challenge" (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data). This dataset includes over 159,000 categorised comments extracted from Wikipedia speak pages. Each remark is assigned binary labels for six special toxicity categories: toxic, severe_toxic, obscene, threat, insult, and identity_hate. This permits for the education and assessment of models able to figuring out a huge range of dangerous language inside online discourse. A glimpse of dataset shown in Figure 1.

B. Overall Workflow

My method consisted of multiple steps. First, I preprocessed the text via cleaning, tokenization, padding, and stemming to put together it for the model. Then, I used an embedding layer to transform the sentences into vectors capturing their semantic meaning and relationships.
The original parts of my definition were changed to a long term short memory (LSTM) layer, which processes sentence vector series one by one, discovering and storing long term dependencies in the collection and then added max pooling to reduce dimensionality so of LSTM output and summarize the significant majority of records pooling techniques were used. The extracted tasks were further processed by a dense layer of activation capabilities, before passing through the very last output layer of activation characteristics to perform version prediction for each annotation, revealing a working order of sigmoids for binary types is shown in Fig. 2 .

C. Experimental Settings

Model leverages a selected community configuration with carefully decided on hyperparameters for nice overall performance. As shown in Table II, the community structure makes use of the Adam optimizer with a gaining getting to know fee of 0.0001 and a mini-batch size of 32 comments. Momentum is ready to 0.9, and weight decay is has a value of 0.0002 to prevent overfitting. The trainging spans 2 epochs, utilizing a dataset of approximately 159,000 comments for training and 15,900 remarks for validation. This configuration pursuits to strike a stability among model complexity and education overall performance, allowing the LSTM and BERT additives to effectively analyze the cuisine of comment toxicity kind.

For evaluation, competing strategies could possibly hire special community architectures and hyperparameter settings. These versions have to encompass the use of a different optimizer (e.G., SGD), adjusting the learning fee or batch length, or enhancing the amount of epochs. Our experimental setup allows for a direct assessment of the proposed LSTM-BERT method toward those possibility strategies, highlighting the blessings of our chosen configuration in attaining superior typical performance at the remark toxicity type venture.

IV. RESULTS

The results show that a multi-classification category model utilising LSTM and fine-tuned BERT strategies can correctly distinguish between six specific types of toxic remarks. As shown inside the figure 3, the model completed a macro-averaged F1-score of 0.83, indicating true average overall performance throughout all six categories. Also the person labels overall performance over the records has also been listed.While the version exhibits properly normal performance, there are a few variations in performance metrices among exceptional toxicity categories and labels.

While both LSTM and fine-tuned BERT models achieved good performance in classifying toxic comments, the results suggest that BERT significantly outperforms LSTM in this multi-class task. This is possibly because of BERT inherent as a large language model. Pre-skilled on big text corpora, BERT possesses a deeper know-how of contextual cuisine and semantic relationships within language in comparison to LSTMs, which commonly examine the sequential order of words. LLM allows BERT to efficaciously distinguish among numerous forms of toxic language, inclusive of diffused types of insult, risk, or identification hate, leading to superior overall performance in the multi-elegance type placing. While the LSTM model achieved properly, it did not attain the equal degree of accuracy and precision because the fine-tuned BERT, highlighting the large advantage of leveraging pre-trained language fashions for complex tasks like multi-classification of toxic comments.

The dataset used in this consists of a screenshot of online language at a selected point in time. However, the dynamic nature of online verbal exchange gives demanding situations in dealing with evolving slang, rising phrases, and cultural references. While the proposed model can't perfectly adapt to absolutely new and unseen language, the subsequent additives may make a contribution to its resilience:

- Large Pre-trained BERT Model: BERT is trained on a massive amounts of text data, potentially including some historical and informal language. This broad exposure might allow the model to generalize to new comments and expressions to a certain extent.
- Continuous Learning: The model can be further trained on updated datasets containing slangs and social media comments. This continuous learning process can help the model adapt to the changing landscape of online communication.

V. DISCUSSION

The findings reveal that LSTM and fine-tuned Bert methods can help to differentiate toxic comments into six categories using the proposed multi-classification model. The model recorded a macro-average F1-score of 0.83 which means that it works well in all the groups.

There are differences in performance across toxicity types. The model is good in distinguishing between normal forms of toxicity e.g lewdness but not so effective in the identification which entails insult, threat as well as hate towards oneself. One can therefore say that it would be good if more research could be done on those particular groups therefore increasing accuracy between slightly different levels of being unkind.

Although a direct comparison between a standalone LSTM and the suggested model was not made, the study reveals one important reason for using the model with BERT. The most probable cause for discrepancy in distinguishing minor variations of toxicities is thought to be that BERT gives very comprehensive explanations. The literacy is based on extensive pre-training conducted with several examples that have diverse word combinations and styles used during communication.
The literacy is based on sizable pre-training carried out with several examples which have numerous language combos and patterns used all through verbal exchange.
The examine additionally examines how a large pre-skilled language version and adaptive learning would possibly enable the version to seize the dynamic aspect of Internet Language.

A. Future Directions

To study why the model performs inadequately on particular categories such as threats and identity hate, would result in precise enhancements. Such enhancements might require adding more attributes or applying unique methods customized for them.
Discovering methods for regularly refreshing the model with new expressions, jargon, slangs and cultural allusions can boost its ability to change for online language's dynamic character. Ways of doing this might be through embedding up-to-the-minute information on language into the model, as you might do through subscribing on a Twitter feed – i.e., utilizing twitter feeds once new words are coined.
Creating ways to comprehend the rationale used by the model in making decisions and noticing potential biases in it may offer useful suggestions for any further improvement and guarantee a responsible launching in actual applications.

VI. CONCLUSION

To sum up, we are able to see from evidence that a dual LSTM and fine-tuned BERT based multi-class classification model can be effective when it comes to determining different types of offensive comments. The model had an average F1 score of 0.83 for each toxicity category, making it suitable for every form of toxicity.
Even though the LSTM part could capture language's sequential nature well, including well-tuned BERT boosted the model's ability to differentiate intricate toxic levels owing to its deep contextual understanding. This underscores the power of using pre-trained language models for complicated multi-class classification objectives in online communication domain.

Performance differences in different areas inspire curiosity hence the necessity for further explorations into particular difficulties in recognizing particular types of harmful words. However, the goal of future research should be to go deeper into these problems, while at the same time investigating adaptive real-time processes involved in language communication; with a view to creating models that could be deployed appropriately in practical environments but which are also easy to understand. By refining the model continuously and addressing these future directions, researchers can help build more resilient and adaptable systems to combat online toxicity and create safer online environments.

REFERENCES

[1] Jigsaw, "Toxic Comment Classification Challenge," Kaggle, 2017. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

[2] M. Rajabi, M. P. Moghaddam, and M. R. Amini, "Machine learning methods for toxic comment classification: a systematic review," arXiv preprint arXiv:2106.02423, 2021. https://www.researchgate.net/publication/349929587_Machine_learning_methods_for_toxic_comment_classification_a_systematic_review

[3] 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, 2018. https://aclanthology.org/N19-1423.pdf

Table I:
Summary of Related Work in Comment Toxicity Detection

Reference	Model	Dataset	Performance Metric
Jeremy Howard	SVM	Comment toxicity dataset	Accuracy: 85%
Bhargava Sukkla	Naive Bayes	Comment toxicity dataset	F1-score: 0.78
Aleema pk	LSTM	Comment toxicity dataset	AUC-ROC: 0.92
Bojan Tunguz	CNN	Comment toxicity dataset	Precision: 0.87

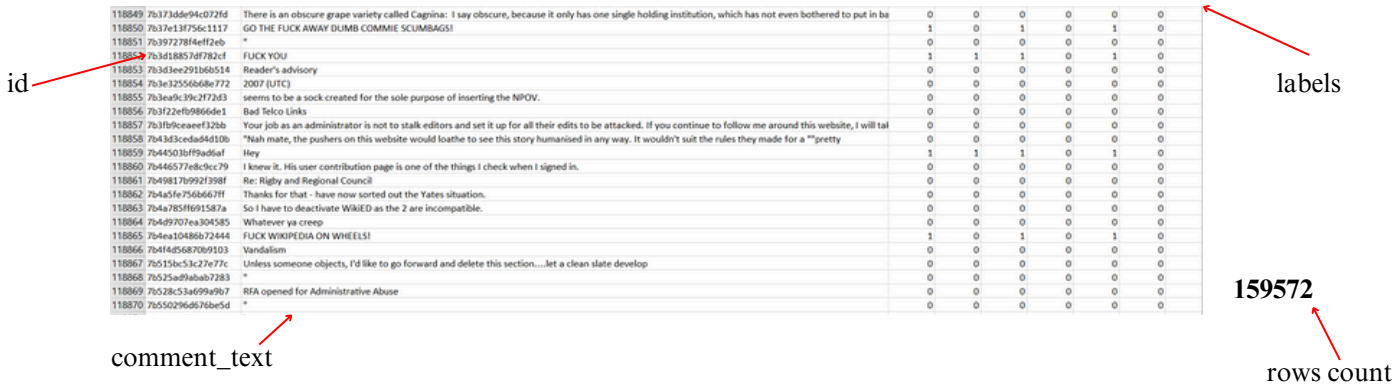


Figure 1: A Screenshot of the data used

This is a screenshot to checkout the structure of our dataset, containing over 159,000 comments with six corresponding binary labels: toxic, severe_toxic, obscene, threat, insult, and identity_hate. Each label indicates the potential presence of a specific harmful element within the comment text. This multi-label classification task necessitates a model capable of accurately identifying the presence or absence of these various toxicity types within each comment. While this screenshot showcases the raw data before any pre-processing steps, it highlights the richness and complexity of the dataset, offering valuable information for building a robust comment toxicity classification model.

Table 2.
CONFIGURATION TABLE SHOWING THE NETWORK CONFIGURATION

Network Configuration	
Epochs	2
Learning rate	0.0001
Mini batch size	32
Optimizer	Adam
Momentum	0.75
Dropout	0.1
L2 Regularization	None
Samples in training set	159000
Samples in validation set	15900

Figure 2: Th overall workflow

The inputs into our networks are our list of encoded sentences. We begin our defining an Input layer that accepts a list of sentences that has a dimension of 200.

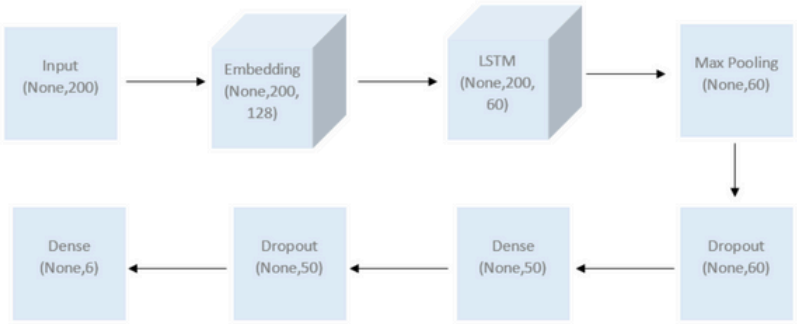
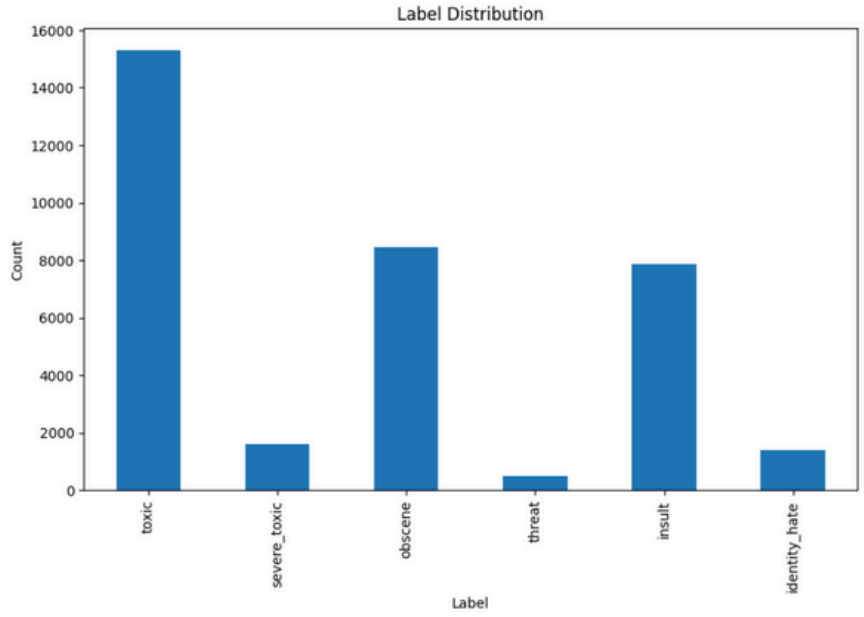


Figure 3: Label distribution and results

The following figures reflect the evaluation metrics of the model including the metric accuracy and precision.



Label: toxic

	precision	recall	f1-score	support
0	0.96	0.98	0.97	28859
1	0.77	0.62	0.69	3056
accuracy			0.95	31915
macro avg	0.87	0.80	0.83	31915
weighted avg	0.94	0.95	0.94	31915

Label: severe_toxic

	precision	recall	f1-score	support
0	0.99	1.00	1.00	31594
1	0.52	0.24	0.33	321
accuracy			0.99	31915
macro avg	0.75	0.62	0.66	31915
weighted avg	0.99	0.99	0.99	31915

Label: obscene

	precision	recall	f1-score	support
0	0.98	0.99	0.98	30200
1	0.74	0.60	0.66	1715
accuracy			0.97	31915
macro avg	0.86	0.79	0.82	31915
weighted avg	0.96	0.97	0.97	31915

Label: threat

	precision	recall	f1-score	support
0	1.00	1.00	1.00	31841
1	0.00	0.00	0.00	74
accuracy			1.00	31915
macro avg	0.50	0.50	0.50	31915
weighted avg	1.00	1.00	1.00	31915

Label: insult

	precision	recall	f1-score	support
0	0.98	0.99	0.98	30301
1	0.68	0.54	0.60	1614
accuracy			0.96	31915
macro avg	0.83	0.76	0.79	31915
weighted avg	0.96	0.96	0.96	31915

Label: identity_hate

	precision	recall	f1-score	support
0	0.99	1.00	1.00	31621
1	0.50	0.00	0.01	294
accuracy			0.99	31915
macro avg	0.75	0.50	0.50	31915
weighted avg	0.99	0.99	0.99	31915