Toxic Comments Classification using Knowledge Distillation

Shadman Ahmad Nafee
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
shadman.ahmad.nafee@g.bracu.ac.bd

Jannatus Sakira Khondaker
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
jannatus.sakira.khondaker@g.bracu.ac.bd

Annajiat Alim Rasel
Computer Science and Engineering
Senior Lecturer, Brac University
Dhaka, Bangladesh
annajiat.alim.rasel@g.bracu.ac.bd

***Abstract*—This research introduces an innovative method for classifying toxic comments by employing knowledge distillation. The approach we employ utilizes a teacher-student framework, in which a substantial DistilBERT model with 66M parameters (teacher) transmits knowledge to a smaller, customized BERT model with 20M parameters (student). The rationale for adopting this technique is the high computing cost involved in implementing the full-scale BERT model, which frequently makes it unfeasible for real-world use cases. The findings of our study indicate that after training, the student model is capable of accurately categorizing offensive remarks, similar to the instructor model, but with significantly reduced computing requirements. This study offers a practical method for utilizing the capabilities of extensive language models in situations when resources are limited.**

***Index Terms*—distillation, DistilBERT, knowledge, BioNER datasets, CNNs, BERT, Naive, LSTM**

I. Introduction

Online platforms have gained significant popularity as a beneficial and widely used medium for individuals to communicate, exchange information, and articulate their viewpoints. Nevertheless, not all online encounters exhibit civility and respect. Some people partake in the practice of posting toxic remarks that exhibit rudeness, animosity, or maliciousness towards others, hence establishing a hostile and disagreeable atmosphere for all individuals involved. Malicious remarks have the potential to improperly impact the psychological and physical well-being, security, and standing of individuals who engage in online activities, as well as the trustworthiness and financial viability of the platforms. Hence, it is crucial to tackle the issue of noxious remarks and devise strategies to diminish or eradicate their occurrence. A significant hurdle in categorizing harmful comments is the creation of models that are both precise and time-effective. Existing methodologies, such as BERT, depend on extensive and intricate pre-trained language models that necessitate substantial quantities of data, computational resources, storage, power, and energy for both training and implementation. These methods are not appropriate for situations with restricted resources or real-time applications, when the availability and cost of resources are limited, and the speed and scalability of the models are crucial. To address this issue, we suggest employing the concept of Knowledge Distillation as a means of diminishing resource utilization. Knowledge Distillation is a machine learning technique that entails transferring knowledge from a larger and more advanced model (referred to as the teacher) to a smaller and less complex model (referred to as the student). The objective is to maintain the efficiency of the initial model while minimizing computing requirements and memory consumption. This is accomplished by instructing the student model to imitate the forecasts or internal representations of the instructor model. Knowledge Distillation is highly beneficial for deploying models on devices with constrained resources. It has been effectively utilized in numerous machine-learning applications, including object detection, speech recognition, and natural language processing. This research presents a new approach for categorizing harmful comments by employing Knowledge Distillation. Our methodology utilizes a teacher-student paradigm, where a substantial DistilBERT model with 66M parameters (teacher) teaches knowledge to a smaller, customized BERT model with 20M parameters (student). The reason for employing this technique is the substantial computational expense associated with implementing the complete BERT model, which frequently renders it unfeasible for real-world applications. The results of our study indicate that after training, the student model can correctly categorize toxic comments, identical to the teacher model, but with substantially fewer computational costs. This study presents a pragmatic approach to harness the possibilities of extensive language models in situations when resources are scarce.

II. Literature Review

Here are key Knowledge Distillation and comment classification advancements. Tahir et al. encourage knowledge distillation. It illustrates how a smart model may improve a simpler model on a single work using its knowledge in several related occupations. He also illuminates knowledge distillation strategies and factors' effects on student models. Mehmood et al. used an MTM teacher and an STM student to condense information. MTM uses many connected BioNER datasets,

while STM uses one. On several BioNER datasets, the authors outperformed baseline models using multiple MTM intermediary layers to compress knowledge to the STM[1]

Quanshi Zhang et al. have demonstrated the significance of a compact student network's ability to learn from a bigger instructor network and achieve exceptional success on this difficult assignment. The research elucidates the efficacy of knowledge distillation in accomplishing classification tasks, by quantifying the knowledge points included in the intermediate layers of deep neural networks (DNNs). The authors put up three assumptions and develop three sets of metrics to test them across a range of classification tasks, such as image classification, 3D point cloud classification, binary sentiment classification, and question answering[2]

Emel Ay suggested transferring information from a voluminous and profound teacher model, namely a fully convolutional network (FCN), to a smaller and less complex student model that has a reduced number of parameters. The student model is taught to imitate the teacher's predictions and feature representations by utilizing a distillation loss function. The studies conducted on many datasets from the UCR time series archive provide encouraging outcomes and indicate that knowledge distillation has the potential to enhance the effectiveness and efficiency of student models for time series classification (TSC).[3]

Ashish et al. identified and categorized toxic internet comments, which are harmful or offensive. Because harmful language is complex and dynamic, addressing it is difficult. Deep learning models like CNNs, RNNs, BERT, and GPT are good at this. However, their efficiency depends on large volumes of data and resources, which may not always be available. Knowledge distillation allows smaller models to learn from bigger models and improve their performance while saving time and money. Knowledge distillation has been applied to various natural language issues, but harmful comment classification continues to be studied. Insufficient diversity in training data leads to biased models that perform poorly on real-world data. Sarcasm, irony, and euphemisms are difficult to recognize and classify[4]

The proposal of Geoffrey et al. suggests utilizing the soft outputs of the instructor network as objectives for the student network, instead of relying on the hard labels of the data. The authors demonstrate that this methodology may enhance the performance and efficiency of the student network across many tasks, including voice recognition and picture classification. Regarding the categorization of toxic comments, it will be a challenge. However, it is possible to train a smaller and more efficient model to imitate a bigger model like mobileBERT. This will decrease the time and cost associated with making predictions, while also enhancing the accessibility and scalability of the classifier[5]

Hao Li et al. presents three models for detecting toxic comments online: Naive Bayes-SVM, LSTM, and BERT. The authors compare their performance on the Civil Comments dataset and use data preprocessing, weighted loss, and ensembling methods to improve the results. The paper shows that BERT achieves the best F1 and EM scores among the models[6]

Jia Cui et al. propose a new way to improve multilingual voice recognition and keyword search models in low-resource languages. Knowledge distillation, which takes knowledge from a larger model (teacher) to a smaller model (student), is suggested to harness the variety and complementarity of multilingual models and features. The authors extensively evaluate Babel OP3 development and surprise languages and find that the student models outperform the teacher and baseline models in word error rate (WER) and maximum term-weighted value. The authors demonstrate that semi-supervised learning and data selection can benefit student models and that the ensemble can learn from one other[7]

Rishabh et al. present a unique approach for distilling semantic knowledge from numerous instructor networks to a student network in federated learning. The approach measures the gap between the probability distributions of the teacher and student models and weights the teacher models based on their intrinsic bias using optimum transport (OT). The research also proposes Semantic Distance (SD) to assess knowledge transmission. The research shows that the proposed technique outperforms entropy-based distillation on SD and is comparable to conventional accuracy and F1 metrics in fine-grained sentiment analysis, conversational emotion identification, and natural language inference[8]

KD approaches for multi-label classification are explained by Youcai et al. Obscene, identity hatred, threat, toxic, insult, severe toxic, and other sorts of toxicity can be found in toxic remarks, making them a typical multi-label text categorization challenge. Thus, utilizing text-based features and classifiers instead of image-based ones, the paper's findings and methodologies may apply to this situation. Text attention maps or word embeddings might be used as dark knowledge for distillation, and the teacher's projected probability could be used to re-weight toxicity activation maps. Another option is to employ transformer-based teacher and student networks like BERT or MobileBERT and test CAMs-based distillation[9]

Data geometry, optimization bias, and strong monotonicity help distillation succeed, according to Mary et al. explains how distillation works and benefits classifier knowledge transmission. Mary et al. propose data geometry and monotonicity-based transfer set and active learning methodologies. The linear distillation model may be used to explain the nonlinear scenario, which is more typical in natural language processing jobs[10]

Felix et al. compares deep learning and machine learning models for this task with the most frequent datasets and assessment measures used in past studies. The report lists Long Short-Term Memory (LSTM) as the most popular and accurate deep learning model and Wikipedia's talk page modifications as the most utilized dataset[11]

Hanwen et al. use data augmentation, model selection, and fine-tuning to develop lightweight models for Chinese NLP applications via knowledge distillation. Hanwen et al. present a lightweight model assessment technique using downstream tasks such named entity identification, keyword extraction, referent deletion, and machine reading comprehension[12]

III. Methodology

A. Algorithm Description:

Knowledgeknowledge distillation is a machine learning technique that involves transferring information from a larger and more sophisticated model (known as the teacher) to a smaller and simpler model (known as the student). The objective is to uphold the performance of the initial model while diminishing computational demands and memory use. This is accomplished by training the student model to imitate the predictions or internal representations of the instructor model. Knowledge distillation is very beneficial for implementing models on devices with limited resources. It has been effectively utilized in many machine-learning applications, such as object identification, audio modeling, and natural language processing.

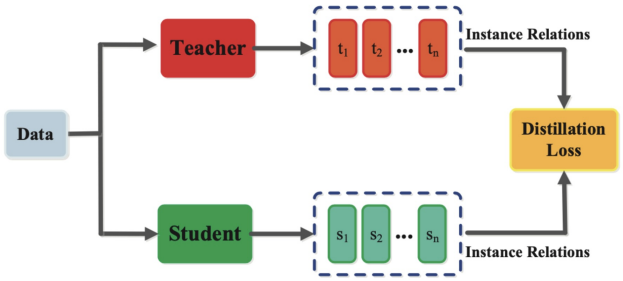

Fig. 1. Knowledge Distillation Flow Chart

B. Data Collection and Data Analysis:

Our dataset consists of a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of toxicity are: toxic, severe_toxic, obscene, threat, insult, identity_hate. We annotated the 'comment_text' as input and as output we showed 0 or 1. Number of data in our dataset is 159571.

C. Data Preprocessing and cleaning:

For data cleaning, we have used the function 'text_preprocessing'. In terms of tokenization, we used 'preprocessing_for_bert'. As a tokenizer we have used 'DistilBertTokenizer'. Our tokenized inputs consist of 'input_ids' and 'attention_mask'.

D. Libraries:

- PyTorch (Maximum code)
- Pandas
- NumPy
- Sklearn/Scikit-Learn
- Matplotlib

E. Models

Our parent model is the DistilBERT, which has 66M parameters. Even though DistilBERT is a faster and smaller version of BERT, it is still a large model and difficult to train. It is not suitable for memoryconstrained devices and takes long inference time. To solve this problem, we have used a custom built BERT model which has 21M parameters. It is one of the DistilBERT model and it can be easily deployed into resourceconstrained devices such as mobile phones or embedded systems. Even though the though, the student model is relatively smaller, it still shows amazing results of above 90% accuracy.

IV. Results

Results achieved by our teacher model-

Accuracy	94.45%
Precision	0.9328
Recall	0.9560
F1 score	0.9442
AUC	0.99

TABLE I
Teacher Model's Result

Results achieved by our student model-

Accuracy	91%
Precision	0.8737
Recall	0.9550
F1	0.9125
AUC	0.9756

TABLE II
Student Model's Result

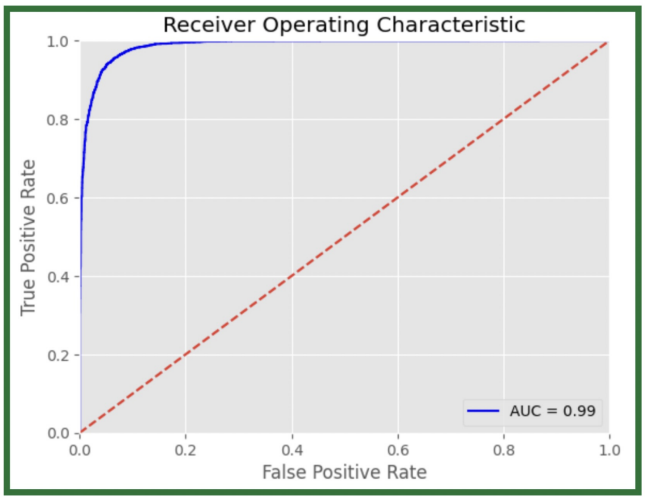

Fig. 2. Teacher Model's Result

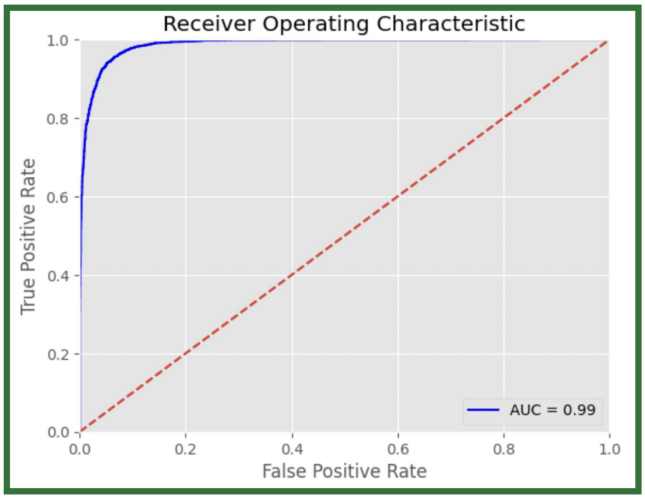

Fig. 3. Student Model's Result

Result	Teacher's Model	Student's Model
Accuracy	94.45%	91%
Precison	0.9328	0.8737
Recall	0.9560	0.9550
F1 Score	0.9442	0.9125
AUC	0.99	0.9756

TABLE III
RESULT COMPARISON TABLE

V. LIMITATIONS

1) The choice of teacher model was supposed to be the BERT model but due to limited resources we had to take DistilBERT as our teacher model.
2) The generalization and scalability of the approach to other types of time series data, such as multivariate, irregular, or streaming data. The analysis of the impact of knowledge distillation on the interpretability and robustness of the student model, as well as the trade-off.
3) Teacher selection has a crucial role in determining the accuracy of the student model. However, it is important to note that the teacher with the greatest accuracy may not necessarily be the most suitable option for distillation.
4) Optimization Challenge: The student model may have difficulty in accurately replicating the prediction distributions of the teacher's model, although having the ability to do so.
5) Generalization trade-off: The student model may enhance its capacity to generalize on the source language or task, but may compromise its performance on other languages or tasks in zero-shot scenarios.

VI. FUTURE WORK

There are a few intriguing potential fields that future research in this field may pursue. First off, other large-scale language models like GPT-3 or RoBERTa, which might have different advantages and disadvantages in terms of performance and computational efficiency, could be added to the teacher-student framework. Second, there is still room for improvement in the distillation procedure. To improve the performance of the student model or further reduce its size, for example, new methods could be created to transfer knowledge from the teacher model to the student model more successfully. Thirdly, the student model's application could go beyond just categorizing harmful comments. Other natural language processing tasks like named entity recognition, sentiment analysis, and question answering could make use of it. Finally, one could look into how the smaller student model affects its interpretability. In addition to being simpler to implement in practical settings, a more condensed and effective model may also be simpler to comprehend and articulate, which is a major benefit in a variety of contexts. Building on the results of the current study, these future research avenues would push the limits of what is feasible with large-scale language models and resource-constrained scenarios. The creation of hybrid models, which combine the advantages of several large-scale language models, is another possible research direction. For example, the combination of DistilBERT, GPT-3, and RoBERTa as teachers could be used to train a student model. This might lead to the creation of a model that outperforms all teacher models combined. Examining various distillation methods could be an intriguing additional avenue to pursue. Although a simple teacher-student framework is used in this study, alternative approaches like self-distillation, multi-teacher distillation, or even ensemble distillation could be investigated. By using these strategies, students may become more effective role models or even surpass the performance of the teacher model in terms of student performance. Moreover, it might be investigated how the distillation procedure affects the student model's robustness. It would be intriguing to look

into whether, in comparison to the teacher model, the distillation process makes the student model more or less vulnerable to adversarial attacks. The use of these models in applications where security is a concern may be significantly impacted by this. Lastly, one could investigate the ethical implications of using simplified models in actual situations. Although the student model's lower processing needs make it more practical for deployment, it is also important to consider the potential for misuse of these models, particularly when it comes to the classification of toxic comments. Future research could look into ways to reduce these risks, like creating procedures for identifying and stopping model abuse. These future lines of inquiry not only expand on the results of the current study but also provide new avenues for the use of large-scale language models in scenarios with constrained resources. They offer stimulating prospects for deepening our comprehension of knowledge extraction and its possible implications for natural language processing.

VII. Conclusion

The research presents a novel approach to categorize harmful comments by employing knowledge distillation, a mechanism that converts information from a big and intricate instructor model to a smaller and more straightforward student model1. The paper use DistilBERT as the teacher model and a bespoke BERT model as the student model. The study demonstrates that the student model may get similar or superior performance than the instructor model across many benchmark datasets, while demanding much less computer resources and inference time. The research additionally contrasts the student model with other conventional machine learning and deep learning models and showcases its superiority in terms of accuracy and robustness. The study presents a pragmatic approach for harnessing the capabilities of extensive language models in situations where there are constraints on resources. The research also proposes potential avenues for enhancing the distillation process, extending the application of the student model to additional tasks in natural language processing, and investigating the ethical ramifications of employing simplified models in real-world scenarios.

VIII. Reference

1) Mehmood, T., Gerevini, A., Lavelli, A., Olivato, M., and Serina, I. (2023). Distilling Knowledge with a Teacher's Multitask Model for Biomedical Named Entity Recognition. Information, 14(5), 255. https://doi.org/10.3390/info14050255
2) Zhang, Q. (2022, August 18). Quantifying the knowledge in a DNN to explain knowledge distillation for classification. arXiv.org. https://arxiv.org/abs/2208.08741
3) A study of Knowledge Distillation in Fully Convolutional Network for Time Series Classification. (2022, July 18). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/document/9892915
4) A comparative study and analysis on toxic comment classification. (2023, June 14). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10169771
5) Hinton, G. (2015, March 9). Distilling the knowledge in a neural network. arXiv.org. https://arxiv.org/abs/1503.02531
6) Li,H., Mao,W., Liu,H. Toxic Comment Detection and Classification. Stanford University Press. https://cs229.stanford.edu/proj2019spr/report/71.pdf
7) Knowledge distillation across ensembles of multilingual models for low-resource languages. (2017, March 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/7953073
8) Bhardwaj, R. (2021, October 6). KNOT: Knowledge Distillation using Optimal Transport for Solving NLP Tasks. arXiv.org. https://arxiv.org/abs/2110.02432
9) Zhang, Y., Qin, Y., Liu, H., Zhang, Y., Li, Y., Gu, X. (2023). Knowledge Distillation from Single to Multi Labels: an Empirical Study. Fudan University Press. https://arxiv.org/pdf/2303.08360.pdf
10) Phuong, M., Lampert, C.H. (2021). Towards Understanding Knowledge Distillation. https://arxiv.org/pdf/2105.13093.pdf
11) Systematic Literature Review: Toxic Comment Classification. (2022, November 4). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9971338
12) Knowledge distillation application technology for Chinese NLP. (2021, January 22). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9362719