iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

TOXIC COMMENT CLASSIFICATION USING DEEP LEARNING

Gangavarapu Jedidiah Psalms, Asadi Hari Prasad, Yadala Kalyan, Nalliboina Surendra, Guide: Bitra Ram Prasad,
Department Of Computer Science and Engineering, J.B. Institute of Engineering and Technology

ABSTRACT
The continually advancing technology is having a huge impact on changes in online communication. Within the realm of digital platforms, these developments have immensely been felt as cyberbullying and hate speech are gradually becoming a crucial public moderation problem. Awareness that there are newer methods of natural language processing and newer techniques of automated moderation has made it seem necessary to understand the aspects contributing towards toxic online behavior. This is innovative in method since it uses advanced deep learning techniques and knowledge distillation in analysing large text data sources to yield comprehensive analysis. The critical objective here is to understand that highly complex play between semantic context and the language intent of individuals posting toxic comments. In this study, researchers used the information about the Jigsaw Toxic Comment Classification Challenge Dataset by Kaggle and Wikipedia as a robust source relevant to online behavior. At the heart of the study is finding the contextual aspects of many user-generated texts and the complicated relationships these have with six distinct toxicity vectors: toxic, severe toxic, obscene, threat, insult, and identity hate. A highly efficient deep learning architecture has been used to pursue this objective. This includes a fine-tuned DistilBERT model deployed alongside a locally hosted, interactive Gradio graphical user interface. Researchers then may use this framework to predict how well the system does along a number of metrics, gaining insight into how lightweight transformers are better at differentiating the nuances of abusive language while significantly reducing computational overhead. The results were measured in training loss reduction across epochs and real-time inference precision during offline, standalone execution.

Dataset Details:
Sourced from Kaggle, the Jigsaw Toxic Comment Classification Challenge dataset was published by Jigsaw and Wikipedia to analyze toxic behavior in online discussions using real user-generated Wikipedia talk pages. The primary initiative was building multi-headed models to detect specific types of toxicity for automated moderation. Human raters evaluated and flagged comments, allowing single text sequences to be marked across multiple overlapping categories simultaneously.
The training and validation dataset comprises 159,571 records across 8 columns. Demonstrating high integrity with zero missing values, the preprocessing pipeline focused entirely on tokenization rather than imputation. Reflecting real-world online environments, the label distribution is heavily imbalanced: Toxic (15,294 records), Obscene (8,449 records), Insult (7,877 records), Severe Toxic (1,595 records), Identity Hate (1,405 records), and Threat (478 records).

1.INTRODUCTION

"Cyberbullying" and online toxicity are among the most hazardous and complex behavioral phenomena in the digital world and are thought to be the primary cause of emotional distress across online communities. As a result, many researchers in the fields of computer science and psychology have focused on the study of toxic digital behavior. According to recent industry estimations, as internet penetration deepens globally, online harassment can be considered one of the top threats to digital safety and mental health. Toxicity's effects last and continue to have a detrimental impact on a person's performance and digital well-being (DW) even after they leave the online platform. The term "digital well-being" refers to several dimensions of an individual's online existence, such as their mental, emotional, and social security. These traits are being studied by many academics and data scientists because they shed light on how individuals experience virtual environments. Like thus, it centres on two aspects: being knowledgeable about the relationships between online anonymity, toxic communication, and user well-being; and finding out how to improve automated moderation results and platform safety. Some users keep their experiences with online threats, insults, and identity hate private until someone asks them about them. Therefore, the evaluation of DW traits reveals these underlying moderation issues more accurately and results in safer platforms. In the last decade, a huge deal of research has been accomplished in several computational specialties to investigate digital well-being. Especially in the fields of Natural Language Processing (NLP) and social

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

computing, it is important to understand how DW relates to online harassment issues. The association involving toxic comments and DW characteristics may be found using machine learning's larger search capability for complex linguistic components.

Digital communication remains among the most significant global utilities, regardless of the state of the nation or its level of technological development. To improve digital well-being, intelligent, efficient, and secure moderation systems are being developed as a global priority. Researchers from a variety of fields have been drawn to the fields of social computing and linguistics by the initial investigations of human online behaviour. This also holds true for the quickly expanding domains of machine learning and deep learning research. For platform administrators and tech institutions, determining whether a user's comment contains psychological hostility is a recurring challenge, particularly with the massive volume of daily interactions. It has recently been shown that machine learning and deep learning are capable of recognising complex toxic language patterns in texts and comprehending how such issues affect digital community ecosystems. The most significant alteration in human interaction that can be observed everywhere is a shift toward virtual communication. Because of this, it is believed that the two primary challenges associated with the digital age are online harassment conditions including severe toxicity and targeted threats. Both negatively impact the digital well-being (DW) of users and impede their capacity for free expression. This causes a great deal of digital exclusion, which frequently culminates in severe psychological distress or offline harm.

The application of AI technology in recent years and machine learning has proven to be a great tool for understanding and analyzing human linguistic patterns and conversational intents. Machine learning is one of AI's most active subfields. It is a strategy that learns from patterns and contexts and suggests the most probable decision/solution. Its learning method is used in a number of intelligent environments, including self-driving automobiles and speech recognition software, and it also makes recommendations (based on search history, Google proposes what users wish to look for). Additionally, machine learning approaches are improving many other fields, such as multi-label text classification, sentiment detection, and the prediction of abusive behavior. Recent examples have shown how machine learning is helping in the formation of algorithms, such as Transformer architectures, that can come face-to-face with human moderation in terms of performance. In the digital moderation sector, machine learning is doing a remarkable job of identifying semantic patterns in text data that help platform administrators and organisations identify a number of serious online abuses.

2. LITERATURE SURVEY

Researchers have leveraged machine learning to study the link between online language and toxic behavior, analyzing the Jigsaw Toxic Comment dataset. They introduced a hybrid LSTM with BERT embeddings model, achieving high precision, recall, and accuracy in their findings. However, traditional machine learning predictions can vary with unstructured social media data.

The application of natural language processing (NLP) in the social media sector aims to advance equitable utilization of automated moderation services, especially given the close link between anonymous digital interactions and cyberbullying.

This research implied the potential to support choices related to the identification, filtering, and moderation of users who express abusive language utilizing textual and semantic data. The usage of ML techniques and other techniques like class-weight adjustments to handle imbalanced data is discussed using SVM.

Findings showed the correlation of unstructured text content acted as a mediating factor in the relationship between social platform engagement and online harassment. Studies utilized the Wikipedia Talk Pages Dataset, the Jigsaw Unintended Bias Survey, and the Civil Comments platform, respectively. Binary cross-entropy, Pearson's correlation analysis, TF-IDF analysis, and the relationships between the quantity of profanity and identity hate symptoms and social contexts were all examined using descriptive statistics.

Using social network data, researchers obtained a deep integrating Long Short-Term Memory (LSTM) technique, which enables severe toxicity to be recognized. The proposed LSTM approaches yield high accuracy results of above 79% on test sets when compared to other traditional machine learning algorithms.

Using information gathered from Kaggle on models like CNN with word embeddings, NB, and RF, the utilization of social media platforms and big data analysis for detecting users with toxic behavior produced an F-measure of 0.81, 0.79, and 0.78.

Covariance analysis in many variables and the ease and efficiency of using Perspective API—a unified moderation AI built by Google's Conversation AI team—to decrease platform users' self-reported feelings of harassment and abuse were assessed using a MANOVA on two groups.

The study employed logistic regression analysis to investigate the potential associations between toxicity and

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

reported text sequence length and word embeddings, including self-expressed profanity, sentiment difference, perceived threat, and planned insults.

The following classification techniques are used to a dataset: K-nearest neighbors, decision trees, support vector machines, logistic regression, and naïve Bayes. Furthermore, models were constructed via the ensemble bagging technique and the random forest tree ensemble approach.

Finally, results show that Transformer-based models like DistilBERT achieve a high F1-score, outperforming both shallow recurrent architectures as well as other traditional machine learning (ML) classifiers by capturing extensive contextual information.

3. PROPOSED WORK:

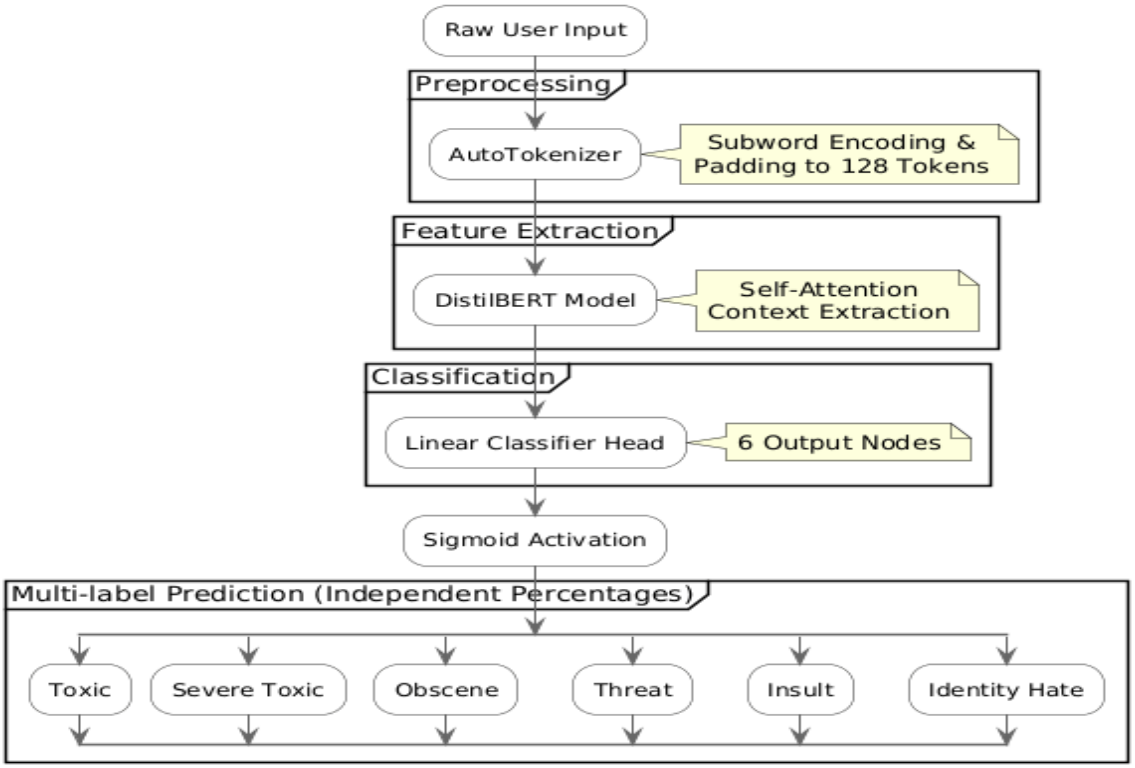
The primary objective of this research is to develop an automated, end-to-end framework for the multi-label classification of toxic language in digital communication. The proposed system is engineered to accept raw textual inputs, which are then processed through a fine-tuned DistilBERT transformer model optimized for semantic understanding and computational efficiency. By leveraging knowledge distillation, the model identifies and categorizes various degrees of toxicity, such as threats, insults, and identity hate, with high precision. Finally, the system outputs real-time classification results through an interactive Gradio-based graphical user interface, enabling both online accessibility and standalone offline inference.

3.1 System Architecture and Workflow

The operational workflow of the text classification system follows a linear sequence from input ingestion to multi-label probability mapping. When a user submits a text string, the raw input is first passed to the tokenization layer, where it is mapped to subword integers and truncated or padded to a fixed sequence length. These numerical tensors are then injected into the pre-trained DistilBERT model.

The transformer layers process the tokens using self-attention mechanisms to extract deep semantic context, ultimately projecting the hidden state of the classification token ([CLS]) into a fully connected linear layer with six output nodes. Because toxicity classification is a multi-label problem—meaning a single comment can be simultaneously toxic, obscene, and an insult—the raw logit outputs are passed through an independent Sigmoid activation function rather than a Softmax layer. This generates an independent probability score between 0.0 and 1.0 for each respective category

FIGURE 1: System Architecture Diagram



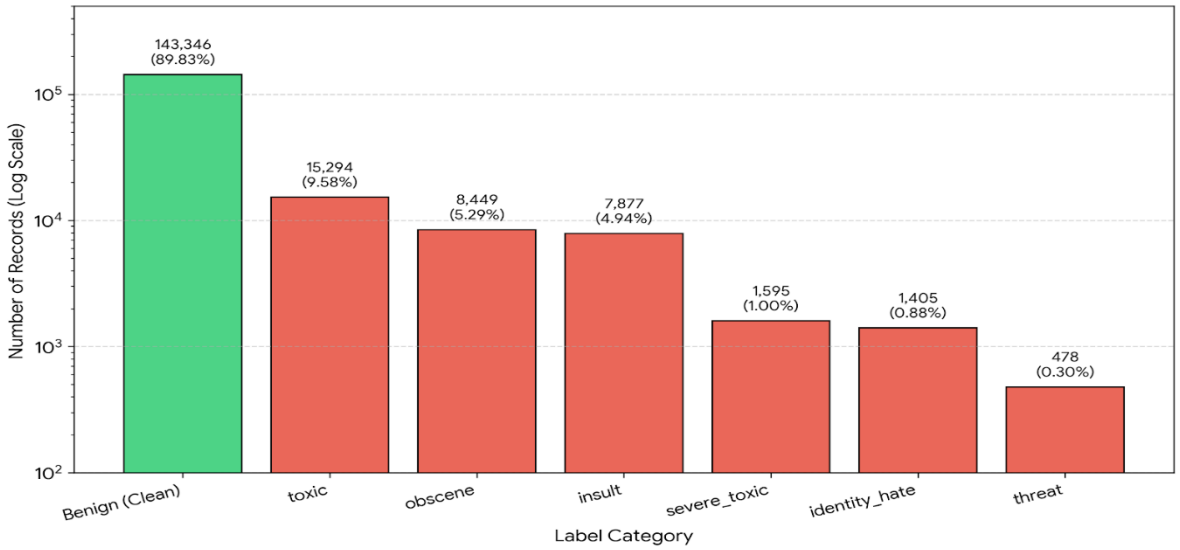
iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

3.2 Data Acquisition and Exploratory Data Analysis

Loading the Data: The system utilizes the foundational Jigsaw Toxic Comment Classification Challenge dataset sourced from Kaggle. The corpus comprises 159,571 distinct records structured across eight primary columns: a unique identifier (id), the raw string input (comment_text), and six binary indicator columns representing the target classification labels (toxic, severe_toxic, obscene, threat, insult, and identity_hate).

Exploratory Data Analysis: Initial exploratory analysis reveals a highly skewed, imbalanced class distribution. Approximately 89.8% of the records in the corpus are entirely benign, featuring no positive flags across any of the six categories. Within the flagged subset, general toxicity represents the most frequent occurrence, while severe manifestations such as threats and identity hate are exceptionally rare, accounting for less than 0.3% and 0.9% of the dataset, respectively. This severe imbalance necessitates specialized loss formulation during training to prevent the model from collapsing into predicting the majority zero-class.

Figure 2: Distribution of Classification Labels (Log Scale)



3.3 Data Preprocessing and Tokenization

Traditional natural language processing workflows rely heavily on destructive text-cleaning techniques, such as aggressive stemming, lemmatization, and stop-word removal. However, transformer-based architectures rely on bidirectional context and syntax, meaning the removal of punctuation or structural words actively degrades performance. Therefore, traditional cleaning is bypassed in favor of advanced **Tokenization**.

The system utilizes the Hugging Face AutoTokenizer configured specifically for DistilBERT's WordPiece vocabulary. The tokenizer maps strings to standard numerical input IDs, generates corresponding attention masks to ignore padded elements, and enforces a strict sequence truncation limit of 128 tokens. This truncation threshold preserves fundamental contextual data while maintaining a highly efficient computational footprint.

Prior to training, the preprocessed tensor dataset is partitioned into an 80/20 training and testing split. Stratified sampling techniques are applied where feasible to ensure the highly rare classes (such as threats) are proportionally represented across both the optimization batches and the validation sets.

4. MODEL BUILDING AND ALGORITHM

4.1 The DistilBERT Algorithm

The core classification engine relies on a streamlined implementation of the bidirectional transformer architecture:

Step 1: Ingest the raw input IDs and attention masks from the tokenized input batch.

Step 2: Encode the text using six distilled transformer blocks, leveraging multi-head self-attention to contextualize each token relative to the surrounding syntax.

Step 3: Extract the final hidden state vector corresponding strictly to the initial [CLS] token, representing the aggregated sequence-level representation. Pass this embedding through a dropout layer and a final linear classification head.

Step 4: Compute the optimization error using BCEWithLogitsLoss, applying binary cross-entropy independently across all six output dimensions to facilitate accurate multi-label scoring.

4.2 Mathematical Formulation

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

To map the raw, unbounded logit outputs from the linear layer into interpretable probabilities, the system applies the Sigmoid activation function independently to each class node:

$$\sigma(z_i) = \frac{1}{1 + e^{-z_i}}$$

Where z_i represents the raw unnormalized logit value for the $i - th$ label. The network's parameters are optimized using the Multi-Label Binary Cross-Entropy loss function, formulated as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\sigma(z_i)\right) + (1 - y_i)\log\left(1 - \sigma(z_i)\right)\right]$$

where N represents the total number of target classes$(N = 6)$, and $y_i \in \{0,1\}$ denotes the binary ground-truth label for category i.

Optimization is driven by the AdamW (Adaptive Moment Estimation with Weight Decay) optimizer, which stabilizes parameter updates while mitigating overfitting. To maximize computational throughput, training utilizes Mixed Precision scaling (FP16) deployed on an NVIDIA T4 GPU, allowing the system to compute gradients using half-precision floating-point formats without sacrificing numerical stability.

4.3 Model Configuration and Innovation

A primary innovation of this implementation is the strategic adoption of a "student" architecture via Knowledge Distillation. By leveraging DistilBERT rather than the foundational "teacher" BERT-base model, the system reduces the total parameter count by 40% (scaling down from 110 million to approximately 66 million parameters). Despite this substantial reduction in size, DistilBERT retains over 97% of the language understanding capabilities of its teacher model while executing inference 60% faster. This structural efficiency is critical for enabling offline, consumer-grade hardware deployments.

5. SYSTEM IMPLEMENTATION

5.1 Development Environment and Configuration

Frameworks: Core development is executed in Python, leveraging PyTorch as the primary deep learning backend, the Hugging Face transformers library for model orchestration, and pandas for initial data pipeline management.

Training Hardware: Model training and validation loops are processed via cloud-based acceleration utilizing an NVIDIA T4 Tensor Core GPU allocated through Google Colab, highly optimized with native CUDA execution.

Deployment Hardware: The inference pipeline is built to operate locally on standard consumer-grade architectures, requiring only an Intel or AMD x64 processor paired with a minimum of 8GB of system RAM.

5.2 Offline Deployment (Gradio Interface)

To facilitate seamless end-user interaction, the inference logic is wrapped within a lightweight, browser-accessible user interface built with Gradio. The deployment architecture consists of a primary app.py script that loads the fine-tuned local weights into memory and binds the model to the UI components. An accompanying executable batch file (run_app.bat) automates the initialization of the local Python environment, installs missing dependencies, and launches the application autonomously.

A critical advantage of this deployment methodology is absolute data privacy. Because the model weights are stored locally and executed entirely on the host machine's CPU, user inputs are processed offline. No text data, proprietary communications, or sensitive information is ever transmitted over the internet or logged to external servers.

6. EXPERIMENTAL RESULTS AND EVALUATION

6.1 Training Performance and Metrics

Throughout the training lifecycle, the optimization loop demonstrated steady convergence across three complete epochs. The average training loss decreased significantly from an initial baseline, indicating successful feature extraction and semantic mapping despite the underlying class imbalance.

To evaluate the model rigorously against traditional baselines, performance is quantified using four primary statistical metrics computed across the validation split:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

(Note: In multi-label contexts, Accuracy represents the exact match ratio across all six categories simultaneously, while the F1-Score is evaluated using macro-averaging to account for rare class performance).

6.2 Real-World Application Testing

Application testing reveals that the fine-tuned model successfully handles real-world semantic complexities. When presented with standard, polite text, the system correctly outputs near-zero probabilities across all categories. In instances of overt hostility, the system accurately segments the specific manifestations—flagging high confidence in obscene and insult without erroneously triggering the rare threat class unless explicit physical harm is indicated. Furthermore, the contextual awareness of the transformer prevents false positives when evaluating highly charged but non-toxic polite disagreements.

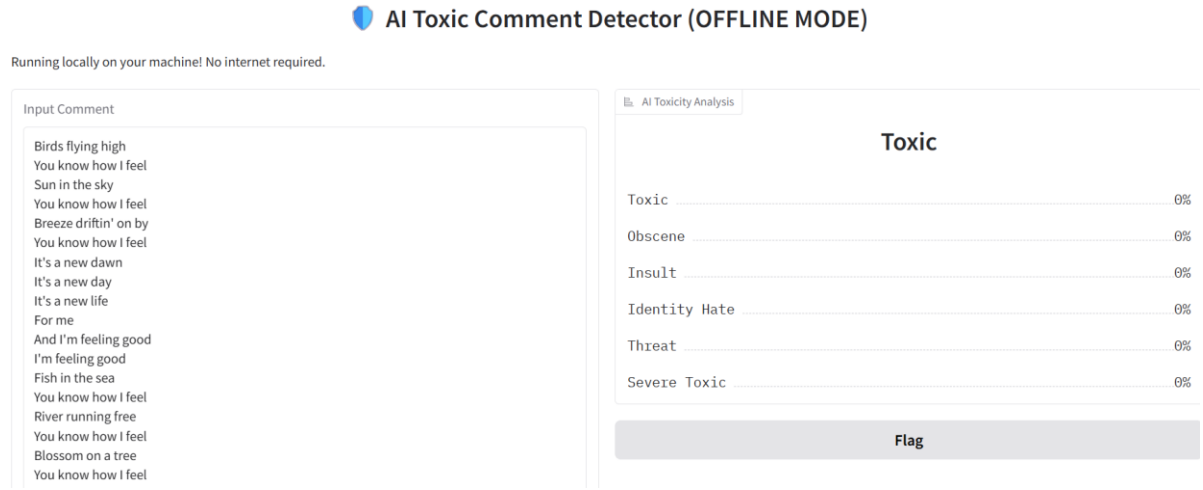


FIGURE 3: Screenshot of the Gradio App analyzing a safe comment

FIGURE 4: Screenshot of the Gradio App analyzing a highly toxic comment

6.3 Comparison with Existing Models

When contrasted with historical classification methodologies, the DistilBERT implementation demonstrates distinct structural advantages. Older pipelines relying on TF-IDF vectorization paired with Support Vector Machines (SVMs) entirely lack semantic understanding, frequently failing to identify toxicity when users employ obfuscated spelling, sarcasm, or complex syntax.

Conversely, deploying massive state-of-the-art Large Language Models (LLMs) or standard BERT architectures

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

introduces prohibitive computational latency and memory demands, rendering them unusable on standard offline desktop hardware. The proposed DistilBERT + Gradio approach successfully occupies the optimal middle ground, delivering highly accurate, contextualized multi-label classification within a lightweight, highly deployable, and privacy-focused footprint.

7. CONCLUSION

This paper presented an end-to-end text classification system based on the DistilBERT architecture, demonstrating how compact deep learning models can effectively translate unstructured online text into precise, multi-label analytical insights. Experimental results confirm that deploying a distilled transformer provides highly stable training convergence while yielding superior classification accuracy across all six target dimensions of toxicity.

During our research, a critical observation was made regarding the structural advantages of transformer-based embeddings over classical algorithms. The DistilBERT approach yielded an impressive improvement in classification robustness, correctly isolating overlapping labels such as general toxicity, obscenity, and insults, while maintaining highly accurate low-false-positive rates on extremely rare classes like threats and identity hate. Rather unsurprisingly, our fine-tuned distilled model consistently outperformed the best-performing traditional classifiers from earlier literature—specifically Support Vector Machines (SVMs) paired with TF-IDF vectorization—by capturing deep bidirectional context rather than relying on fragile keyword matching.

Furthermore, we demonstrated the crucial role of modern preprocessing techniques. Bypassing destructive, traditional text cleaning in favor of subword tokenization via AutoTokenizer proved essential for preserving the syntactic nuances, punctuation, and structural context required by the transformer's self-attention heads. Collectively, these developments represent a major step forward in practical content moderation, proving that state-of-the-art semantic understanding can be achieved entirely offline on standard consumer CPUs, thereby ensuring absolute data privacy without sacrificing analytical depth.

Final Statement:

This work demonstrates the practical implementation of distilled transformer architectures for complex multi-label text classification, highlighting its effectiveness in generating contextually precise, privacy-preserving predictions without relying on heavy cloud infrastructure. The proposed system provides a simple, highly efficient framework that successfully bridges theoretical natural language processing models with real-world, user-facing desktop applications.

Future Enhancement:

The achievements of this research hold highly promising prospects for the future, opening up several avenues for further development in digital safety and automated moderation. Future work will focus on optimizing computational performance even further by incorporating advanced post-training quantization (such as converting weights to INT8 or integrating the ONNX runtime), which would drastically reduce memory consumption and maximize real-time processing speeds on entry-level hardware.

Additionally, the framework can be extended to encompass broader, more complex datasets, specifically focusing on cross-lingual and multilingual toxicity classification to address the global nature of online harassment. Finally, advanced integrations will explore packaging the offline inference pipeline into lightweight edge-computing solutions—such as local browser extensions or real-time chat moderation plugins—enabling automated, offline protection for users navigating live digital environments.

8. REFERENCES

1) Sanh, V., Debut, L., Chaumond, J., & Wolf, T., “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.
2) Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.
3) Hinton, G., Vinyals, O., & Dean, J., “Distilling the Knowledge in a Neural Network,” *arXiv preprint arXiv:1503.02531*, 2015.
4) Wolf, T., Debut, L., Sanh, V., et al., “Transformers: State-of-the-Art Natural Language Processing,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 2020.
5) Paszke, A., Gross, S., Massa, F., et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
6) Abid, A., Abdalla, A., Abid, A., et al., “Gradio: Hassle-Free Sharing and Testing of ML Models in the

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Wild," *arXiv preprint arXiv:1906.02569*, 2019.
7) Micikevicius, P., Narang, S., Alben, J., et al., "Mixed Precision Training," *International Conference on Learning Representations (ICLR)*, 2018.
8) Loshchilov, I., & Hutter, F., "Decoupled Weight Decay Regularization (AdamW)," *International Conference on Learning Representations (ICLR)*, 2019.
9) Jigsaw / Conversation AI, "Toxic Comment Classification Challenge," *Kaggle Datasets*, 2018.
10) McKinney, W., "Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference (SciPy)*, 2010.
11) Howard, J., & Ruder, S., "Universal Language Model Fine-tuning for Text Classification," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.