Plagiarism Scan Report

4%
Plagiarism

2%
Exact Match

2%
Partial Match

96%
Unique

Words	809
Characters	6313
Sentences	62
Paragraphs	55
Read Time	5 minute(s)
Speak Time	6 minute(s)

Content Checked For Plagiarism

ABSTRACT

This project introduces a deep learning-based system for multi-format toxicity detection across text, audio, and images. Toxic online interactions—such as hate speech, threats, and harassment—pose significant risks to user safety and digital community health. Manual moderation and rule-based methods are insufficient for addressing nuanced or context-dependent toxic behavior. To tackle these challenges, the project leverages a transformer-based model, specifically BERT, fine-tuned for multi-label classification. This model processes complex language patterns to detect subtle toxicity often missed by traditional methods.

The system expands toxicity detection to spoken formats and textual content in images through advanced techniques like speech-to-text transcription and OCR. With rigorous training on a comprehensive dataset, the model achieves high precision and recall in detecting categories such as hate speech, threats, and identity-based abuse. By addressing multiple input types and enabling real-time content moderation, the project aims to promote safer online interactions while preserving freedom of expression.

INTRODUCTION

The rise of social media and online communication has amplified both positive engagement and harmful interactions. Toxic online comments—ranging from hate speech to harassment—threaten individual well-being and constructive discourse. This project focuses on creating an Advanced Toxicity Classification System using natural language processing (NLP) and deep learning to detect toxic content across diverse inputs like text, audio, and images. The system's goal is to support online platforms in fostering healthier digital environments.

OBJECTIVE

The project aims to develop a high-performing toxicity detection system leveraging the BERT model for multi-label classification. This includes accurately identifying different toxicity types—hate speech, threats, obscenity, and identity attacks—within user-generated content.

SCOPE OF THE PROJECT

The system addresses multi-format toxicity detection, focusing on detecting various abusive behaviors in text, spoken content, and textual data extracted from images. It is designed for integration into social media platforms, forums, and other digital spaces.

FEATURE EXTRACTION

Feature extraction is crucial for processing raw data into structured information usable by machine learning models. This project incorporates different techniques for text, audio, and images:

Text-Based Features:

Using BERT, contextual embeddings are generated for words in sentences. Key steps include:

Tokenization: Text is split into subwords with special tokens ([CLS], [SEP]).
Contextual Embeddings: BERT captures bidirectional context, improving language understanding.
Audio Features:

Speech-to-Text Conversion: The Google Speech Recognition API transcribes audio into text, which is then processed similarly to text inputs.
Sentiment Analysis: An additional layer captures tonal nuances like sarcasm or aggression.
Image Features:

Optical Character Recognition (OCR): Extracts text from images for further analysis.
Text Embeddings: Extracted text is processed through BERT for toxicity detection.
Contextual Text Processing:

Slang Normalization: Expands abbreviations and slang (e.g., "lol" → "laughing out loud").
Spelling Correction: Corrects errors using tools like TextBlob.
LITERATURE SURVEY
Recent studies highlight challenges and advancements in toxicity detection:

ToxicChat (2023): Explores toxicity detection in user-AI conversations, emphasizing domain-specific datasets.
ToxiSpanSE (2023): Focuses on domain-specific toxicity in code reviews, advocating for explainability in AI models.
Prompt-Based Detection (2022): Utilizes generative approaches with large language models, showing sensitivity to prompt design.
Bias in Detection Models (2021): Investigates biases in toxicity models, stressing fairness in subgroup performance.
Bias Mitigation with Deep Learning (2023): Balances detection accuracy with bias reduction using transformer models like BERT.
METHODOLOGY
1. Data Collection:
The project uses the Jigsaw Toxic Comment Classification dataset, which contains approximately 150,000 labeled samples across categories like toxic, severe toxic, obscene, threat, insult, and identity hate.

2. Data Preprocessing:

Text cleaning (removing special characters, URLs).
Tokenization with the BERT tokenizer to convert text into input IDs and attention masks.
Normalization (lowercasing, expanding contractions).
3. Model Architecture:

The system employs BERT, fine-tuned for multi-label classification, adding custom layers for toxicity categories.
Hyperparameters include a learning rate of 2e-5, batch size of 16, and three training epochs.
4. Training and Testing:

Split: 80% training, 20% validation.
Optimizer: AdamW with cross-entropy loss for multi-label classification.
Early stopping prevents overfitting.
5. Evaluation Metrics:

Precision and Recall: Measure true positive and false positive rates.
F1-Score: Balances precision and recall.
ROC-AUC: Evaluates model performance across classes.

IMPLEMENTATION

The system integrates into a web application using the Gradio library, offering an interactive interface for toxicity detection. Users can input text, audio, or images, which are processed through respective feature extraction pipelines (e.g., speech-to-text, OCR). The BERT-based model then provides toxicity predictions.

RESULTS

The model demonstrates high accuracy and strong performance in multi-label classification, with precision and recall improving over traditional approaches. Confusion matrices and ROC curves illustrate its effectiveness across categories. Challenges include high computational demands and handling subtle or implicit toxicity, such as sarcasm.

CONCLUSION

This project successfully developed a toxicity detection model leveraging BERT's transformer-based architecture. The model enhances content moderation systems by identifying nuanced toxic language, surpassing traditional methods in accuracy and adaptability. While the results are promising, challenges remain in improving computational efficiency and addressing biases in training data. Further research is needed to enhance the model's scalability and robustness against evolving toxic language patterns.

Matched Source

Similarity 6%

Title:12.5 Social media content analysis

Aug 21, 2024 — This method combines quantitative and qualitative approaches to analyze text, images, and networks. Researchers must navigate ethical concerns, ...

https://library.fiveable.me/communication-research-methods/unit-12/social-media-content-analysis/study-guide/hU7sy3MuBl4CaXZ5

Similarity 3%

Title:Precision and recall

Recall in this context is also referred to as the true positive rate or sensitivity, and precision is also referred to as positive predictive value (PPV); ...

https://en.wikipedia.org/wiki/Precision_and_recall