

Toxic Chat Detector

A Machine Learning Approach to Identify Toxicity in
Online Gaming



Omar Husain (100491847)

CSCI 4050U - Machine Learning Final Project

Problem Statement

-  Online gaming communities suffer from a widespread toxicity problem, including insults, harassment, and hate speech.
-  This negative behavior degrades the gaming experience and can drive players away from the community.
-  Our solution is an automated system that uses machine learning to detect and flag toxic chat messages in real-time, promoting a healthier online environment.

Learning Problem Definition

The task is framed as a supervised Binary Text Classification problem.



The model takes a raw text string as input and outputs a binary prediction indicating whether the message is toxic or not.

Dataset Overview

Custom Gaming Dataset

~2,200 labeled samples

TOXIC EXAMPLES:

"you are trash", "uninstall noob", "worst player ever"

NON-TOXIC EXAMPLES:

"great game", "gg well played", "nice shot"

Twitter Hate Speech

~25,000 labeled samples

A larger, more general dataset used for comparison and to evaluate model generalization on a different domain of text.

Model 1: Baseline

TF-IDF + Logistic Regression

Raw Text



TF-IDF Vectorizer



Logistic Regression

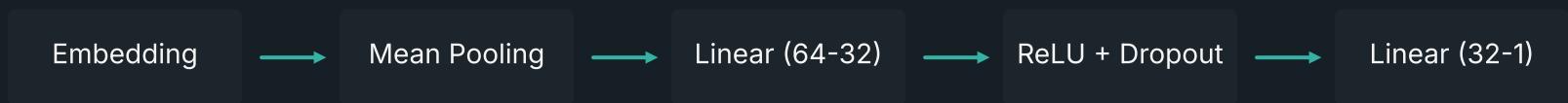
- **TF-IDF:** Converts text to numerical features.
- **max_features:** 5000
- **ngram_range:** (1, 2) - uses words & word pairs
- **class_weight:** 'balanced' - handles data imbalance



Model 2: Neural Network

PyTorch Sequential Model

A simple but effective neural network for text classification.



Total Parameters: ~350,000

Embedding Dimension: 64

Dropout Rate: 0.3

Training Process

Baseline Model



Training Time

~5 Seconds

Neural Network



Training Time

~2 Minutes

Results Comparison

Metric	Baseline	Neural Network
Accuracy	94%	92%
Precision	95%	93%
Recall	94%	91%
F1 Score	94%	91%

Key Findings



The simpler TF-IDF baseline outperformed the neural network.

Why did this happen?

- ➊ The dataset is relatively small (~2,200 samples).
- ➋ TF-IDF effectively captures important keywords and phrases (n-grams) which are strong indicators of toxicity in this domain.
- ➌ Neural networks typically require much larger datasets to learn complex patterns and outperform traditional ML methods.

Future Work



Use Pre-trained Models

Leverage large language models like BERT for potentially higher accuracy.



Multi-Label Classification

Detect specific types of toxicity (e.g., threat, insult, identity hate).



Multi-Language Support

Expand the model to detect toxicity in languages other than English.



API Integration

Provide an API for real game developers to integrate the service.



Questions?

Thank you for your attention.