Conversational Toxicity Detection

Emanuele Fontana
Università degli Studi di Bari Aldo Moro
e.fontana7@studenti.uniba.it

Problem Statement and Objectives

Key Challenges

- Online platforms harbor toxic interactions
- Limited work on Italian language toxicity
- Need for real-time detection capabilities

Main Objectives

Developing systems for:

- Toxic conversation detection
- Personality classification
- Real-time toxicity detection

Dataset Construction Pipeline

Existing Toxic Dataset IDaToC:

- Annotated Italian conversations
- Various toxicity types
- Emotional manipulation
- Psychological violence

Generated Non-Toxic Dataset:

- Google Gemini API
- Healthy conversations
- Corpus balancing

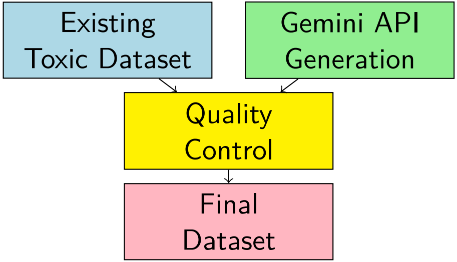

Figure: Dataset generation pipeline

Dataset Generation Parameters

- **Model**: Gemini-2.0-flash-lite - Fast inference with quality generation
- **Temperature**: 1.8 - High creativity for diverse conversation styles
- **Top-p**: 0.95 - Nucleus sampling for coherent text generation
- **Top-k**: 40 - Limits selection to the 40 most probable tokens at each generation step
- **Max tokens**: 2048 - Maximum conversation length per generation

Overall Approach

Three Main Components

1. **Binary Classification**: Traditional ML baseline
2. **Personality Classification**: Zero-shot + Fine-tuning
3. **Real-Time Detection**: Personality-based system

BERT Model

`BERT-base-italian-xxl-cased` with personality tokens

Binary Classification

Compared Approaches:

- **Approach 1**: Raw text + TF-IDF
- **Approach 2**: Italian preprocessing + TF-IDF

Italian Preprocessing Pipeline:

- SpaCy (it_core_news_sm)
- Lemmatization
- Stop words removal
- Italian-specific tokenization

Model Configuration:

- Logistic Regression
- Nested Cross-Validation (5-fold)
- Hyperparameter grid search

Personality Classification with BERT

BERT Model Enhancement

- Base model: `dbmdz/bert-base-italian-xxl-cased`
- Added personality tokens: `[NARCISISTA]`, `[MANIPOLATORE]`, etc.

Two Approaches Comparison

Zero-Shot Learning:

- No training required
- Similarity-based classification
- Uses personality descriptions
- Cosine similarity matching

Fine-Tuning:

- Task-specific training
- Custom classifier head

Real-Time Detection System

Detection Mechanism

- Message-by-message analysis
- Context-aware predictions
- Weighted confidence scoring
- Adaptive threshold: 0.3
- Immediate toxicity alerts

Weighted Scoring Formula

$$\text{toxic\_score} = \sum_{i=1}^{n} w_i \times \text{confidence}_i \tag{1}$$

$$\text{avg\_score} = \frac{\text{toxic\_score}}{n} \tag{2}$$

$$\text{is\_toxic} = \text{avg\_score} > 0.3 \tag{3}$$

Binary Classification Results

Table: Binary Classification Performance

Approach	Accuracy	F1	Precision	Recall
Raw Text	1.0000	1.0000	1.0000	1.0000
Preprocessed	1.0000	1.0000	1.0000	1.0000

Important Insight

Preprocessing requires 20x more computational time without performance benefits

Personality Classification - Zero-Shot vs Fine-Tuned

Table: Zero-Shot Performance

Metric	Score
Accuracy	0.0268
Macro Precision	0.0010
Macro Recall	0.0364
Macro F1-Score	0.0020

Table: Fine-Tuned Performance

Metric	Score
Accuracy	0.5628
Macro Precision	0.5093
Macro Recall	0.5043
Macro F1-Score	0.5015

Performance Improvement:

- Zero-shot: 2.68%
- Fine-tuned: 56.28%
- **21x improvement!**

Confusion Matrix for Personality Classification

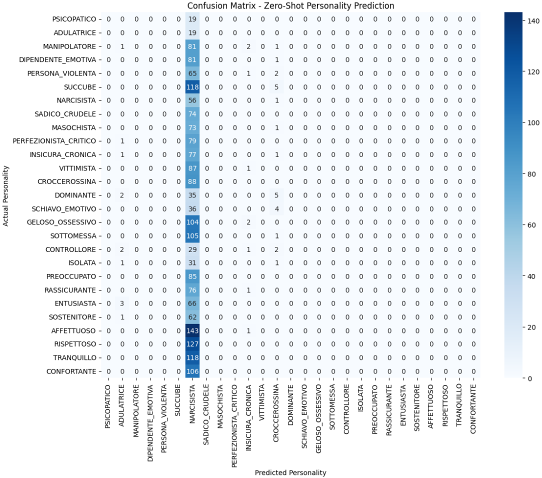

Figure: Zero-Shot Classification

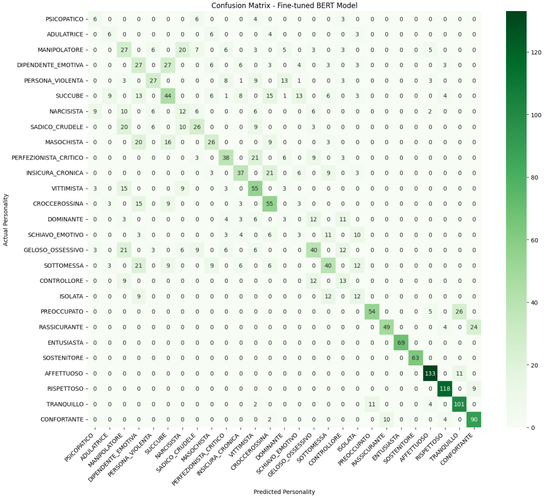

Figure: Fine-Tuned Classification

Real-Time Toxicity Detection

Table: Real-Time System Performance

Metric	Score
Accuracy	0.9884
Precision	0.9943
Recall	0.8889
F1-Score	0.9915

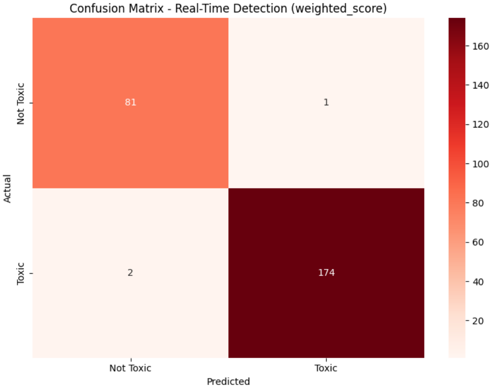

Figure: Confusion Matrix for Real-Time Detection

Main Contributions

Key Results

- **Binary Classification**: Perfect performance without preprocessing
- **Personality**: Fine-tuning significantly outperforms zero-shot
- **Real-Time**: 98.84% accuracy in real-time toxicity detection

Innovations

- Automatic pipeline for non-toxic data generation
- Integration of personality classification + toxicity detection
- Adaptive system with weighted scoring

Limitations and Future Work

Current Limitations

- Specific to Italian language
- Limited personality framework
- Limited context window (512 tokens)

Future Directions

- Multilingual extension
- GPT-based architectures

Availability

Code and dataset available on GitHub:
`https://github.com/Fonty02/NLP/tree/main/Exam`

Thank You for Your Attention

Thank You! 🚀