Conversational Toxicity Detection and Real-Time Toxicity Assessment using BERT-based Models

A Personality Classification Approach

Emanuele Fontana

Università degli Studi di Bari Aldo Moro

Outline

1. Introduction and Motivations
2. Dataset
3. Methodology
4. Experimental Results
5. Conclusions

Problem and Motivations

- Online conversation platforms often contain toxic interactions
- Traditional approaches focus on individual messages or keywords
- Need to capture complex conversational dynamics
- Focus on Italian conversations with psychologically abusive behaviors

Work Objectives

Main Objective

Develop a comprehensive system that combines:

- Personality classification (28 types)
- Real-time toxicity detection
- Conversational context analysis

Contributions

- BERT-based system for Italian conversations
- Synthetic non-toxic data generation
- Hybrid approach: zero-shot + fine-tuning

Dataset Construction

Existing Toxic Dataset:

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

Generated Non-Toxic Dataset:

- Google Gemini API
- Healthy conversations
- Positive dynamics
- Corpus balancing

Generation Pipeline

- Temperature: 1.8 for variety
- Format validation with regex
- Multi-level quality control
- Integration and standardization

Overall Approach

Three Main Components

1. **Binary Classification**: Traditional Machine Learning
2. **Personality Classification**: BERT zero-shot + fine-tuning
3. **Real-Time Detection**: System based on personality patterns

BERT Model Used

`BERT-base-italian-xxl-cased`

- Specialized for Italian language
- 28 personality types
- Context window of 512 tokens

Binary Classification - Baseline

Compared Approaches:

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

Italian Preprocessing:

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

Hyperparameter Tuning:

- Logistic Regression
- C: [0.01, 0.1, 1, 10]
- Penalty: ['l1', 'l2']
- 5-fold cross-validation

Personality Classification

Zero-Shot Approach

- Pre-trained BERT embeddings
- Detailed personality descriptions
- Cosine similarity for matching
- Incremental context construction

Fine-Tuned Approach

- Dropout regularization (0.3)
- AdamW optimizer
- Early stopping on validation loss

Binary Classification Results

Table: Binary Classification Performance

Approach	Accuracy	F1-Score	Precision	Recall
Without Preprocessing	1.0000	1.0000	1.0000	1.0000
With Preprocessing	1.0000	1.0000	1.0000	1.0000

Important Insight

- Identical performance for both approaches
- Preprocessing requires 20x more computational time
- **Recommendation**: Use raw text for efficiency

Personality Classification Results

Zero-Shot:

Metric	Score
Accuracy	0.0268
Macro F1	0.0020

Fine-Tuned:

Metric	Score
Accuracy	0.5628
Macro F1	0.5015

Analysis

- **Significant improvement**: 2.68% 56.28% accuracy
- Zero-shot limited in generalizing personality types
- Fine-tuning captures complex conversational dynamics

Real-Time Detection

Table: Real-Time System Performance

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

Scoring Mechanism

- Message-by-message analysis
- Weighted scoring based on personality
- Alert threshold: 0.3
- Conversational context adaptation

Main Contributions

Key Results

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

Innovations

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

Limitations and Future Work

Current Limitations:

- Specific to Italian language
- 28 personality framework
- Limited context window (512 tokens)
- Domain-specific dataset

Future Directions:

- Multilingual extension
- Larger datasets
- GPT-based architectures
- Real-world deployment
- Extended context windows

Availability

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

Thank You for Your Attention

Emanuele Fontana
e.fontana7@studenti.uniba.it

Università degli Studi di Bari Aldo Moro