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

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

***Abstract*—This paper presents two studies on conversational toxicity. The first study investigates toxic conversation detection using simple Machine Learning models. The second study describes a novel and well-documented BERT-based system for real-time toxic conversation tagging and recognition through fine-tuning, which also performs personality classification in dialogues, incorporating zero-shot learning and fine-tuning approaches. Both studies yielded excellent results, with the proposed methodologies achieving significant performance in their respective tasks and demonstrating the effectiveness of context-aware processing in conversational AI applications.**

## I. Introduction and Motivations

Online conversation platforms and social media have become integral parts of modern communication, yet they often harbor toxic interactions that can cause psychological harm and create hostile environments. The automatic detection of toxic conversations and personality-driven behaviors has emerged as a critical challenge in maintaining healthy digital spaces.

Traditional approaches to toxicity detection often focus on individual messages or keywords, failing to capture the nuanced dynamics of conversational context and personality-driven behaviors. This limitation becomes particularly evident in intimate relationship conversations where toxicity manifests through subtle manipulation, emotional control, and psychological abuse patterns rather than explicit offensive language.

Our work addresses these challenges by developing a comprehensive system that combines personality classification with real-time toxicity detection, leveraging the contextual understanding capabilities of BERT-based models. The system is designed to identify toxic personality patterns in Italian conversational data, providing both immediate detection capabilities and detailed personality analysis.

## II. Related Work

**TODO** Recent advances in transformer-based models have significantly improved natural language understanding tasks, particularly in the domain of conversation analysis. BERT [1] and its variants have shown remarkable performance in various NLP applications, including sentiment analysis and text classification.

Previous work in toxicity detection has primarily focused on single-message classification [2], while conversation-level analysis remains underexplored. Personality detection in text has been addressed through various psychological frameworks [3], but the application to real-time conversational toxicity detection represents a novel contribution.

The integration of zero-shot learning approaches with fine-tuned models for conversation analysis has shown promising results in recent studies [4], providing the foundation for our comparative analysis between different BERT-based approaches.

## III. Proposed Approach

Our comprehensive methodology encompasses four main components: (1) automated dataset generation using Google's Gemini API, (2) binary classification using traditional machine learning approaches, (3) a dual-approach personality classification system using both zero-shot and fine-tuned BERT models, and (4) a real-time toxicity detection system based on personality patterns.

### *A. Dataset Generation and Integration*

Our dataset construction approach combines existing toxic conversation data with newly generated non-toxic conversations to create a balanced training corpus. The toxic conversation dataset was already available from previous work, containing Italian conversational data with detailed toxicity annotations. To complement this, we developed an automated generation pipeline using Google's Gemini API specifically for creating high-quality non-toxic conversational scenarios.

*1) Existing Toxic Dataset Integration:* The foundation of our dataset construction is built upon a pre-existing collection of toxic conversations in Italian. This dataset contains conversational exchanges exhibiting various forms of toxicity including manipulation, emotional abuse, control patterns, and psychological violence. Each conversation in the toxic dataset is accompanied by detailed annotations explaining the specific toxic behaviors and relationship dynamics present.

*2) Non-Toxic Dataset Generation:* To create a balanced corpus, we developed a comprehensive generation pipeline for non-toxic conversations using Google's Gemini API.

This automated approach ensures diversity while maintaining conversational authenticity and cultural appropriateness for Italian speakers.

*3) Generation Architecture and Configuration:* The generation system employs Gemini-2.0-flash-lite as the base model with carefully tuned parameters to ensure conversational diversity and authenticity:

- **Temperature**: 1.8 to encourage creative and varied responses
- **Top-p**: 0.95 for nucleus sampling to balance creativity with coherence
- **Top-k**: 40 to maintain reasonable vocabulary constraints
- **Max Output Tokens**: 2048 to accommodate detailed conversations

Safety settings are configured to block medium and above content for harassment, hate speech, sexually explicit material, and dangerous content, ensuring the generated conversations remain within appropriate boundaries while still capturing toxic behavioral patterns.

*4) Prompt Engineering for Conversational Realism:* The generation process utilizes carefully crafted prompts designed to elicit realistic conversational patterns. For toxic conversations, the prompt structure includes:

1) Specification of toxic relationship dynamics
2) Requirements for numbered, alternating dialogue between two participants
3) Emphasis on Italian language usage and cultural context

For non-toxic conversations, the prompts focus on healthy relationship dynamics such as "Entusiasta e Sostenitore", "Preoccupato e Rassicurante", "Affettuoso e Rispettoso", and "Tranquillo e Confortante".

*5) Quality Control and Validation:* The generation pipeline implements multiple validation layers:

- **Format Validation**: Regex patterns ensure proper conversation structure with numbered, quoted messages
- **Content Completeness**: Verification that all required fields (couple type, names, conversation, explanation) are present
- **Conversation Length**: Ensures minimum message requirements for meaningful interactions
- **Retry Mechanism**: Up to 3 attempts per conversation with exponential backoff for failed generations

*6) Dataset Integration and Final Processing:* The final dataset creation process involves merging the existing toxic conversations with the newly generated non-toxic conversations:

1) **Format Standardization**: Both datasets are processed to ensure consistent conversation format and metadata structure
2) **Label Assignment**: Toxic conversations are labeled with value 1, while generated non-toxic conversations receive label 0
3) **Conversation Format Verification**: Regex-based validation ensures proper message structure with quoted, numbered alternating dialogue
4) **Dataset Concatenation**: Merging of both datasets while preserving conversation integrity and metadata consistency
5) **Final Cleaning**: Removal of conversations that fail validation criteria, ensuring dataset quality and consistency

The integration process produces a balanced corpus containing both authentic toxic relationship patterns from the original dataset and diverse healthy relationship dynamics from the generated conversations, providing a comprehensive foundation for both binary classification and personality-based analysis.

### *B. Binary Classification Approach*

To establish a baseline understanding of the dataset's separability, we implemented a traditional machine learning approach for binary toxicity classification. This study compares the effectiveness of text preprocessing on Italian conversational data.

*1) Feature Extraction and Vectorization:* We employ TF-IDF (Term Frequency-Inverse Document Frequency) vectorization as our primary feature extraction method. Two distinct approaches are evaluated:

**Approach 1: Raw Text Processing**

- Direct application of TF-IDF to unprocessed conversation text
- Minimal computational overhead
- Preservation of original linguistic patterns and colloquialisms

**Approach 2: Italian Language Preprocessing**

- Utilization of spaCy's Italian language model (it_core_news_sm)
- Tokenization with Italian-specific rules
- Lemmatization to reduce words to their base forms
- Stop word removal using Italian stop word lists
- Punctuation and numeric token filtering

*2) Model Architecture and Hyperparameter Optimization:* The classification employs Logistic Regression with comprehensive hyperparameter tuning through nested cross-validation:

**Hyperparameter Grid:**

- **Regularization Strength (C)**: [0.01, 0.1, 1, 10]
- **Penalty Type**: ['l1', 'l2'] for feature selection and ridge regularization
- **Maximum Iterations**: [100, 200, 500] to ensure convergence
- **Solver**: 'liblinear' for compatibility with both L1 and L2 penalties

**Cross-Validation Strategy:**

- Inner loop: 5-fold cross-validation for hyperparameter selection

- Outer loop: 5-fold cross-validation for unbiased performance estimation
- Grid search optimization using accuracy as the primary metric
- Final model evaluation on held-out test set (30% of data)

*3) Computational Efficiency Analysis:* The binary classification study includes a detailed computational cost analysis comparing preprocessing approaches:

- Processing time measurement for text preprocessing pipelines
- Memory usage comparison between raw and processed text representations
- Performance-to-cost ratio evaluation
- Scalability assessment for real-world deployment scenarios

### C. Dataset Preparation and Personality Tagging

For the BERT-based personality classification task, the dataset consists of conversational exchanges between individuals exhibiting various personality types, categorized into toxic and non-toxic relationships. We developed a robust preprocessing pipeline that:

1) Validates conversation format using regex patterns to ensure proper message structure
2) Extracts and cleans individual messages from conversational threads
3) Maps personality couple types to individual personality classifications
4) Creates personality tokens in the format [PERSONALITY] for model training
5) Applies context-aware tagging to maintain conversational coherence

The personality mapping includes 28 distinct personality types, ranging from toxic patterns (e.g., PSICOPATICO, MANIPOLATORE, NARCISISTA) to healthy relationship dynamics (e.g., AFFETTUOSO, RISPETTOSO, RASSICURANTE).

### D. Classification Approaches

### E. BERT-based Personality Classification

We implemented two complementary approaches for personality classification:

*1) Zero-Shot Learning Approach:* The zero-shot method leverages pre-trained BERT embeddings to classify personalities without task-specific training. The process involves:

1) Creating detailed personality descriptions for each of the 28 personality types
2) Computing contextual embeddings for both conversation messages and personality descriptions
3) Using cosine similarity to match messages with the most appropriate personality type
4) Building conversational context incrementally to improve prediction accuracy

*2) Fine-Tuned Classification Model:* The fine-tuning approach adapts BERT specifically for personality classification:

1) Extending the tokenizer vocabulary with personality tokens
2) Implementing a custom PersonalityClassifier with dropout regularization
3) Training with early stopping based on validation loss
4) Incorporating conversational context in the input representation

The fine-tuned model architecture includes: - BERT-base-italian-xxl-cased as the base model - Custom classification head with dropout (rate: 0.3) - AdamW optimizer with linear learning rate scheduling - Maximum sequence length of 512 tokens

### F. Real-Time Toxicity Detection

The real-time detection system analyzes conversations message-by-message, using a weighted scoring mechanism:

1) Predicts personality type for each incoming message using conversational context
2) Calculates toxicity scores based on personality classification confidence
3) Applies weighted scoring where toxic personalities increase the score and healthy personalities decrease it
4) Triggers toxicity alerts when the average weighted score exceeds a threshold (0.3)

## IV. Experimental Setup and Results

### A. Dataset Characteristics

The final dataset comprises 14 different personality couple combinations with a total of approximately 2,000 conversations after cleaning and validation. The dataset is balanced between toxic (labeled as 1) and non-toxic (labeled as 0) conversations.

### B. Binary Classification Baseline

Initial experiments compared TF-IDF vectorization with and without Italian text preprocessing using Logistic Regression:

Table I: Binary Classification Results

| 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 |

Both approaches achieved perfect classification performance, indicating well-separated classes in the dataset. However, preprocessing required 20Œ more computational time without performance benefits.

### C. Personality Classification Results

*1) Zero-Shot Performance:* The zero-shot approach achieved: - Accuracy: 0.2847 - Macro Precision: 0.1032 - Macro Recall: 0.1081 - Macro F1-Score: 0.1045

*2) Fine-Tuned Model Performance:* The fine-tuned model significantly outperformed zero-shot learning: - Accuracy: 0.8235 - Macro Precision: 0.8127 - Macro Recall: 0.8194 - Macro F1-Score: 0.8147

Training converged after 15 epochs with early stopping, achieving a validation loss of 0.4823.

### D. Real-Time Detection Performance

The real-time toxicity detection system using weighted scoring achieved: - Accuracy: 0.8906 - Precision: 0.8924 - Recall: 0.8889 - F1-Score: 0.8906 - Average processing time: 0.089 seconds per conversation

## V. Analysis and Discussion

### A. Model Comparison

The fine-tuned BERT model demonstrated substantial improvements over zero-shot learning, with accuracy increasing from 28.47% to 82.35%. This improvement can be attributed to:

1) Task-specific adaptation through fine-tuning
2) Better handling of conversational context
3) Improved representation learning for personality-specific patterns

### B. Context-Aware Processing

Both approaches incorporated conversational context by: - Maintaining dialogue history during prediction - Building cumulative context for improved accuracy - Using personality-aware conversation reconstruction

This context-aware approach proved crucial for understanding personality dynamics in multi-turn conversations.

### C. Real-Time Detection Effectiveness

The weighted scoring approach for real-time detection provides several advantages: - Immediate toxicity alerts from the first message - Confidence-weighted scoring for nuanced detection - Reduced false positives through healthy personality score reduction

## VI. Limitations and Future Work

### A. Current Limitations

1) **Language Specificity**: The system is currently optimized for Italian conversations
2) **Dataset Size**: Limited to approximately 2,000 conversations
3) **Personality Framework**: Relies on a specific 28-personality classification scheme
4) **Context Window**: Limited by BERT's maximum sequence length (512 tokens)

### B. Future Directions

1) **Multilingual Extension**: Adapting the system for other languages
2) **Larger Datasets**: Expanding training data for improved generalization
3) **Advanced Architectures**: Exploring GPT-based models and longer context windows
4) **Real-world Deployment**: Integration with chat platforms and social media monitoring

## VII. Conclusion

We presented a comprehensive system for conversational personality detection and real-time toxicity assessment using BERT-based models. The system demonstrates significant improvements in personality classification accuracy (82.35%) compared to zero-shot approaches (28.47%) and achieves effective real-time detection performance (89.06

The integration of context-aware processing with personality-based toxicity detection represents a novel contribution to the field of conversational AI safety. The system's ability to process conversations in real-time while maintaining high accuracy makes it suitable for practical deployment in online communication platforms.

Our work establishes a foundation for more sophisticated conversational analysis systems and demonstrates the importance of personality-aware approaches in toxicity detection. The comprehensive evaluation across multiple metrics and approaches provides valuable insights for future research in this domain.

## VIII. Code and Data Availability

The complete implementation, including data preprocessing pipelines, model training scripts, and evaluation frameworks, is available in the project repository. The system includes:

1) Data cleaning and validation scripts
2) Zero-shot and fine-tuned model implementations
3) Real-time detection system with configurable thresholds
4) Comprehensive evaluation metrics and visualization tools

## References


[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[2] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/15004

[3] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf