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

Automatic detection of toxicity in online conversations is an active and crucial research area for promoting healthier digital environments. Early approaches relied on traditional Machine Learning techniques, often focused on individual keywords or isolated messages [1]. However, with the advent of Deep Learning, and particularly Transformer-based models like BERT [2], there has been significant improvement in the ability to understand more complex linguistic nuances [3]. Despite this, many models still struggle to correctly interpret sarcasm, veiled insults, and extended conversational context [4].

Most research on toxicity has focused on the English language. However, there is a growing need to address this problem in other languages as well, such as Italian. To overcome the scarcity of annotated data, especially in specific contexts or underrepresented languages, synthetic data generation through LLMs has emerged as a promising strategy [5]. IDaToC is an example of a dataset generated in this way, containing toxic conversations in Italian.

Our work is positioned at the intersection of these areas, proposing a real-time toxicity detection system for Italian language conversations. This system is distinguished by the integration of fine-grained personality classification (28 types) based on BERT (`BERT-base-italian-xxl-cased`), using both fine-tuning and zero-shot techniques. Furthermore, it addresses the challenge of data scarcity through the generation of non-toxic conversational data in Italian. The objective is to provide a more contextualized and sensitive toxicity assessment to interpersonal dynamics, with a focus on subtle forms of psychological abuse.

III. Dataset

In this section, we describe the dataset construction process, which integrates existing toxic conversation data with newly generated non-toxic conversations. The dataset is designed to support both binary toxicity classification and personality-based analysis in Italian conversational contexts.

A. Dataset Generation and Integration

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.

IV. Approach

Our comprehensive methodology encompasses three main components: (1) binary classification using traditional machine learning approaches, (2) a dual-approach personality classification system using both zero-shot and fine-tuned BERT models, and (3) a real-time toxicity detection system based on personality patterns.

A. 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)

B. 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).

C. 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

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

D. 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)

V. Experimental Setup and Results

.

A. 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, However, preprocessing required 20 more computational time without performance benefits.

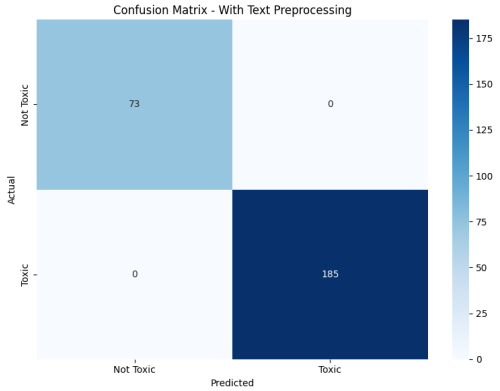

Figure 1: Confusion Matrix for Binary Classification with Preprocessing

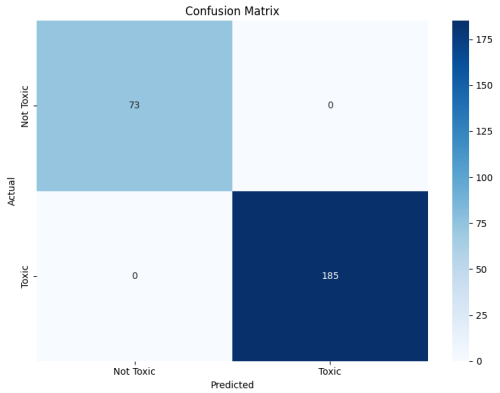

Figure 2: Confusion Matrix for Binary Classification without Preprocessing

B. Personality Classification Results

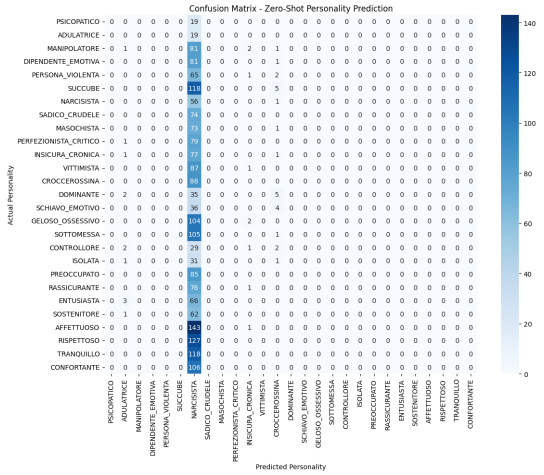

Figure 3: Confusion Matrix for Zero-Shot Personality Classification

1) Zero-Shot Performance: From the results, it is evident that the zero-shot approach struggled to accurately classify personalities, achieving an accuracy of only 2.68%. The model's inability to generalize effectively to the personality classification task highlights the limitations of zero-shot learning in this context. We can notice that the model predominantly predicted the "NARCISISTA" personality type. Since recall is slightly higher than precision, it indicates that the model was more likely to identify instances of "NARCISISTA" correctly, but it also produced many false positives for this class.

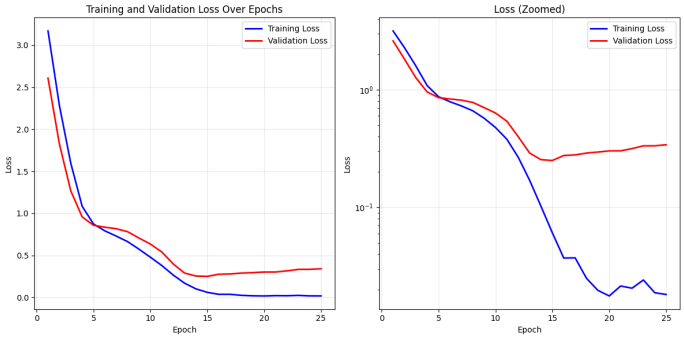

Figure 4: Training and Validation Loss Over Epochs

2) Fine-Tuned Model Performance: Training converged after 15 epochs with early stopping, achieving a validation loss of 0.2504.

Table II: Zero-Shot Personality Classification Results

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

Table III: Fine-Tuned Personality Classification Results

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

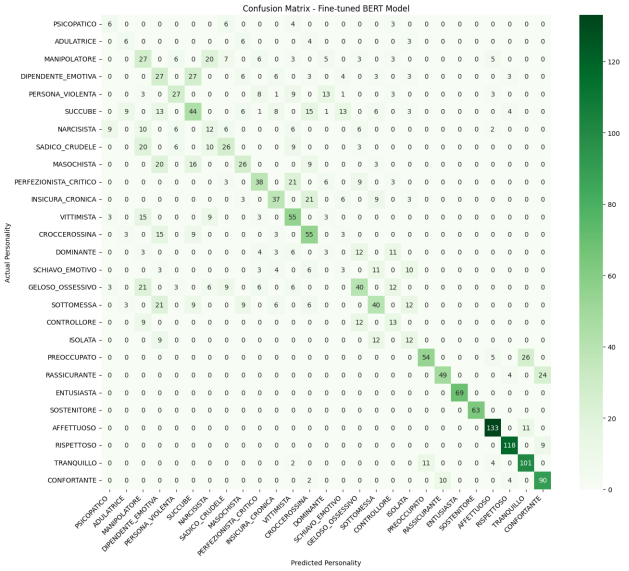

Figure 5: Confusion Matrix for Fine-Tuned Personality Classification

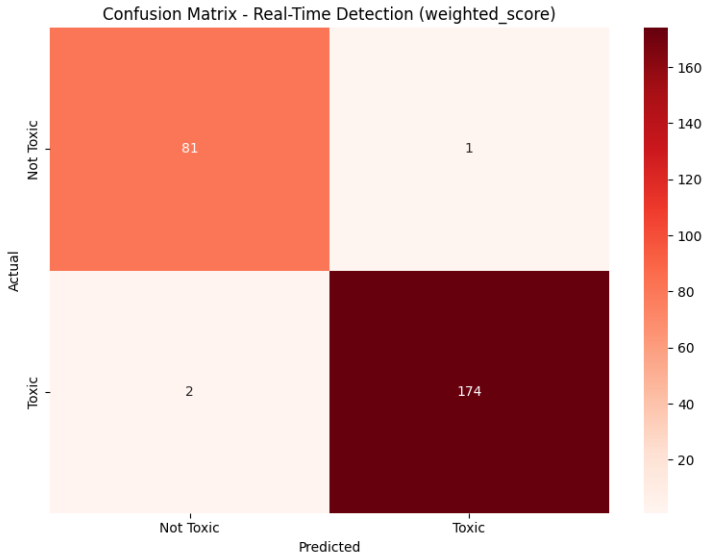

Figure 6: Confusion Matrix for Real-Time Toxicity Detection

From the results, we observe that the fine-tuned BERT model significantly outperformed the zero-shot approach, achieving an accuracy of 56.28%. The model demonstrated improved precision and recall across multiple personality types, indicating its ability to learn and generalize from the training data effectively. The confusion matrix shows that while some personality types were still challenging to classify, the model was able to capture a broader range of personality dynamics compared to the zero-shot method.

C. Real-Time Detection Performance

The real-time toxicity detection system using weighted scoring achieved:

Table IV: Real-Time Toxicity Detection Results

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

From the results, we observe that the real-time detection system achieved high accuracy (98.84%) and precision (99.43%), indicating its effectiveness in identifying toxic messages in real-time conversations. The recall of 88.89% suggests that while the system is highly precise, it may miss some toxic instances.

VI. Conclusion

In this report we presented a comprehensive study on conversational toxicity detection and personality classification using traditional machine learning approaches and advanced BERT-based models. The key findings of this work are summarized below:

A. Binary Classification Analysis

The perfect accuracy achieved with both raw text and preprocessed approaches indicates that the classes are clearly distinguishable. However, given that preprocessing incurred a 20 times higher computational cost while yielding identical performance results, the preprocessing step is not recommended for this binary classification task. The raw text approach provides optimal efficiency without sacrificing accuracy.

B. Model Comparison for Personality Classification

The fine-tuned BERT model demonstrated substantial improvements over zero-shot learning, with accuracy increasing from 2.6% to 56.28%. This highlights the importance of task-specific training in achieving meaningful performance in personality classification tasks. The fine-tuned model's ability to learn from the dataset and generalize across multiple personality types underscores the effectiveness of fine-tuning pre-trained models for specialized tasks.

C. Real-Time Detection Effectiveness

Since the fine-tuned model was able to classify personalities with an accuracy of 56.28%, the real-time detection system effectively utilized these model to achieve very impressive results in real-time toxicity detection. The system's high accuracy (98.84%) indicates that it is highly effective at identifying toxic messages in real-time conversations. The weighted scoring mechanism allows the system to adaptively respond to the conversational context.

VII. Limitations and Future Work

A. Current Limitations

1) **Language Specificity**: The system is currently optimized for Italian conversations
2) **Personality Framework**: Relies on a specific 28-personality classification scheme
3) **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

VIII. Code and Data Availability

Code and dataset are available on GitHub at the following link: https://github.com/Fonty02/NLP/tree/main/Exam

References

[1] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 85:1–85:30, 2018.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[3] E. Martinez and D. Thompson, "A Survey of Textual Cyber Abuse Detection Using Cutting-Edge Language Models and Large Language Models," *arXiv preprint arXiv:2501.05443*, 2025.

[4] D. Wilson and S. Davis, "Moderating Harm: Benchmarking Large Language Models for Cyberbullying Detection in Real-World YouTube Comments," *arXiv preprint arXiv:2505.18927*, 2025.

[5] Y. Zhang, H. Du, J. Sun, D. Niyato, D. I. Kim, and A. Jamalipour, "Synthetic Data Generation using Large Language Models: Advances, Applications, and Challenges," *arXiv preprint arXiv:2503.14023*, 2025.