Toxicity, couple dynamics and most toxic sentence detection in Italian conversations using classical machine learning models and transformer-based models

Davide Cirilli
Università degli Studi di Bari Aldo Moro

Overview

Context and Goals

Why this work matters

- Toxic behaviors in online conversations can cause serious psychological and social harm.
- Italian conversational toxicity and personality-driven patterns are underexplored in NLP research.
- There is a need for systems that go beyond keyword detection to capture complex interaction dynamics.

Our contribution

- A system combining toxicity detection and personality analysis in Italian dialogues.
- Integration of classical ML (e.g., Logistic Regression, Naive Bayes) and transformer models (BERT, BART).
- A web app for interactive real-time couple dynamics and toxic sentence detection.

Preprocessing and Generation Settings

Toxic data preprocessing

- Removed incomplete / malformed entries
- Extracted most toxic sentence via regular expressions
- Normalized whitespace, punctuation, and symbols

Non-toxic data generation and parameters

- **Model**: **gemini-2.0-flash-lite** via **API + Google AI Studio**
- **Generation Settings**:
  - **Temperature**: 1.8
  - **Top-p**: 1.0
  - **Top-k**: 0
  - **Safety filters**: BLOCK_MEDIUM_AND_ABOVE (harassment, hate speech, explicit, dangerous)

Final Dataset Composition

Toxic Conversations

- ~950 Italian dialogues with toxic dynamics
- Includes manipulation, emotional control, psychological abuse
- Annotated with relationship type and most toxic sentence

Non-Toxic Conversations

- ~600 synthetic dialogues created with Gemini API and Google AI Studio
- Healthy, culturally appropriate interactions
- 5 Positive dynamics: *Supportive-Insecure*, *Collaborative-Propositive*, etc.

Binary Toxic Classification: Overview

Goal

Classify conversations as **toxic** or **non-toxic** based on dialogue content.

Modeling strategies

- Traditional ML with TF-IDF + Logistic Regression
- Traditional ML with TF-IDF + Multinomial Naive Bayes
- Experiments with and without linguistic preprocessing

Binary Toxic Classification: Details

Preprocessing variants

- **Without preprocessing:** raw text, tokenized, bi-gram TF-IDF
- **With preprocessing:** spaCy Italian model (lemmatization, stopword and punctuation removal, lowercasing)

Training parameters

- 80-20 stratified train/test split
- Grid search + 5-fold stratified CV
- TF-IDF bi-grams, max_features tuned (2000-7000)
- Logistic Regression: tuned $C \in \{0.1, 1, 10\}$
- Naive Bayes: tuned $\alpha \in \{0.1, 1\}$

Couple Dynamics Prediction: Task and Goal

Objective

Classify the conversational dynamic between two participants into one of 15 predefined relationship types.

Examples of classes

- *Narcissist and Submissive*
- *Manipulator and Emotionally Dependent*
- *Supportive and Insecure*
- *Grateful and Appreciative*

Couple Dynamics Prediction: Modeling Approaches

Explored approaches

- **TF-IDF + LSA + Logistic Regression**
- **Frozen BERT embeddings + Logistic Regression**
- **Fine-tuned BERT**

Pipeline differences

- LSA: dimensionality reduction with TruncatedSVD
- Frozen BERT: [CLS] embeddings without task-specific tuning
- Fine-tuned BERT: direct adaptation to the classification task

BERT model: `dbmdz/bert-base-italian-uncased`

Couple Dynamics Prediction: Training Configurations

TF-IDF + LSA + Logistic Regression

- Unigrams and bigrams TF-IDF
- SVD components: 100, 200, 300 (grid search)
- Logistic Regression: $C \in \{0.1, 1, 10\}$ (lbfgs solver)

Frozen BERT + Logistic Regression

- Extract [CLS] token embeddings
- Logistic Regression: $C = 0.1$, max 2000 iterations

Fine-tuned BERT

- Adam + weight decay, learning rate warmup
- Batch size: 8, max length: 512
- Up to 7 epochs, early stopping (patience $= 2$)
- Evaluation: accuracy, macro precision, recall, F1

Most Toxic Sentence Detection: Task and Goal

Objective

Identify or generate the most toxic sentence within a given conversation.

Challenges

- Toxic phrases can be subtle, context-dependent, or implicit.
- Class imbalance: toxic sentences are rare compared to non-toxic ones.
- Open-ended phrasing makes exact match generation difficult.

Two complementary approaches: classification and generation

Most Toxic Sentence Detection: Classification Approach

Model

Fine-tuned BERT (`dbmdz/bert-base-italian-uncased`)

Pipeline

- Input: individual message + couple dynamic type (as additional context)
- Output: toxicity probability for each message
- Most toxic sentence = message with highest toxicity probability

Training configuration

- Max length: 128 tokens
- Batch size: 32, learning rate: 2×10^{-5}
- AdamW optimizer, weight decay: 0.01
- Up to 7 epochs, early stopping (patience = 2)
- Stratified 80/10/10 split: train/val/test

Most Toxic Sentence Detection: Generation Approach

Model

Fine-tuned BART (`facebook/bart-base`)

Pipeline

- Input: full conversation prefixed with a task descriptor
- Output: generated most toxic sentence
- Generation decoding: beam search (4 beams), early stopping

Training configuration

- Max input length: 512 tokens, output: 64 tokens
- Batch size: 4, learning rate: 3×10^{-5}
- Early stopping (patience $= 2$)
- GroupShuffleSplit to avoid similar conversations across sets

Web Application: Summary

Purpose

- Real-time couple dynamics classification
- Most toxic sentence detection (classification + generation)

Key features

- Chat-like interface, dynamic updates after each input
- On-demand toxic phrase extraction (classification + seq2seq)

Built with Gradio

Binary Toxicity Classification: Performance Summary

Model	Accuracy	Precision	Recall	F1
Logistic Regression (raw)	0.9968	0.9968	0.9968	0.9968
Naive Bayes (raw)	0.9968	0.9968	0.9968	0.9968
Logistic Regression (spaCy)	0.9936	0.9936	0.9936	0.9936
Naive Bayes (spaCy)	0.9968	0.9968	0.9968	0.9968

Table: Binary toxicity classification results

Near-perfect performance across all configurations

Binary Toxicity Classification: Confusion Matrices (Raw Text)

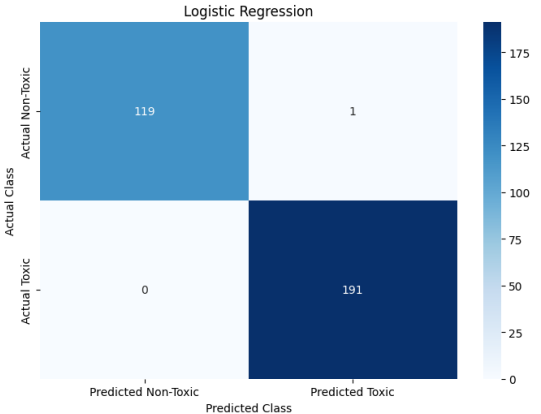

Figure: Logistic Regression (raw text)

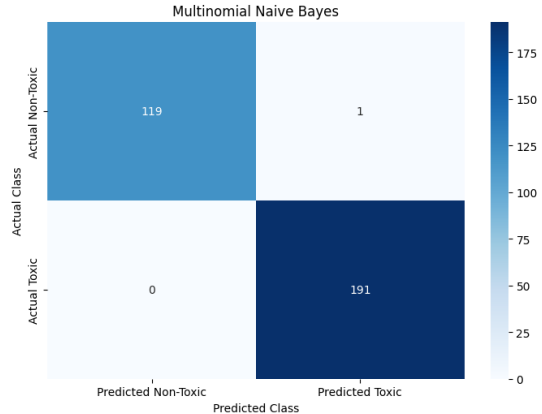

Figure: Naive Bayes (raw text)

Binary Toxicity Classification: Confusion Matrices (spaCy Processed)

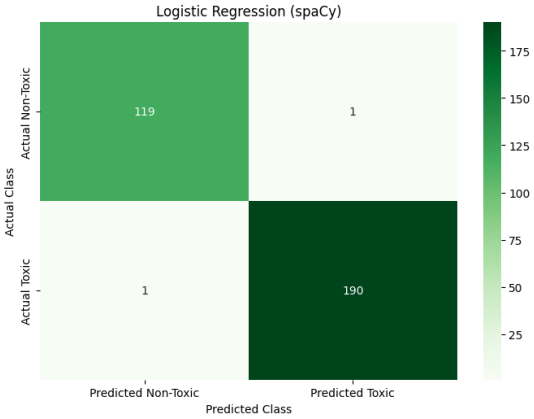

Figure: Logistic Regression (spaCy processed)

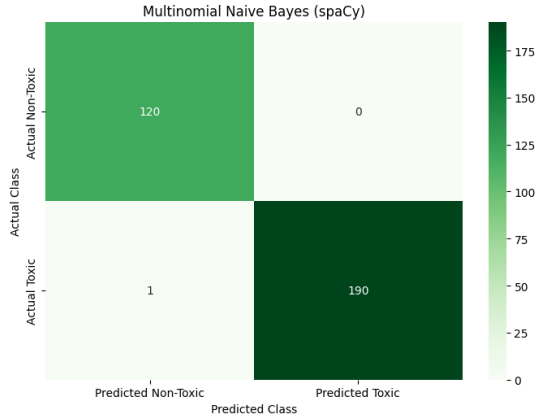

Figure: Naive Bayes (spaCy processed)

Couple Dynamics Prediction: Performance Summary

Approach	Accuracy	Precision	Recall	F1
LSA + Logistic Regression	0.7395	0.7233	0.7160	0.7113
Frozen BERT + Logistic Regression	0.7010	0.6840	0.6767	0.6753
Fine-tuned BERT	0.8013	0.8093	0.7867	0.7780

Table: Couple dynamics classification results (macro average)

Fine-tuned BERT outperforms other approaches

Couple Dynamics Prediction: Confusion Matrices

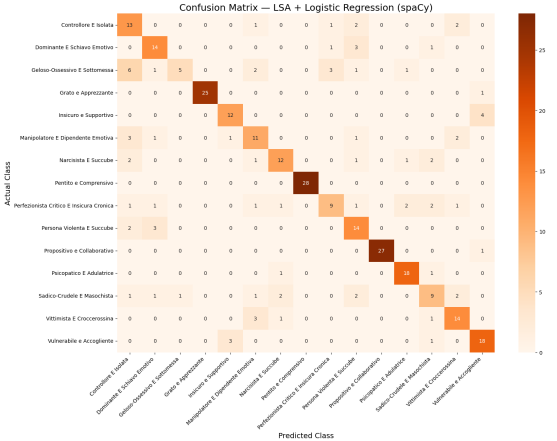

Figure: Confusion matrix - LSA + Logistic Regression

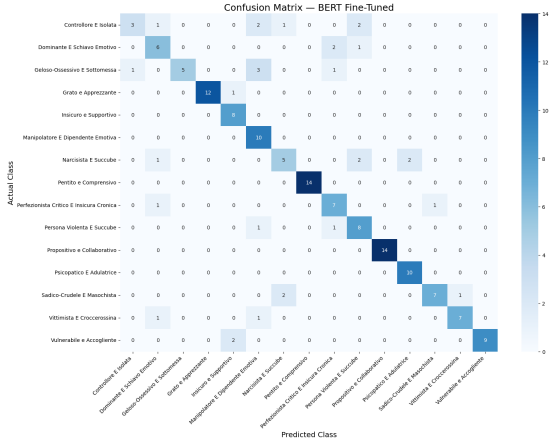

Figure: Confusion matrix - Fine-tuned BERT

Couple Dynamics Prediction: Fine-tuned BERT Loss Curves

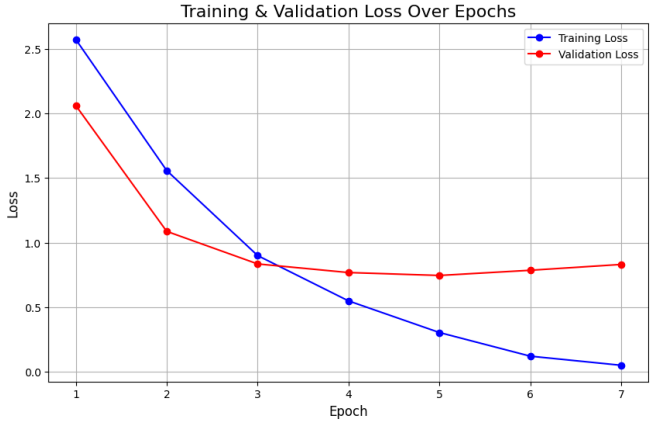

Figure: Training and validation loss during fine-tuning

Good convergence with no significant overfitting observed

Most Toxic Sentence Detection: Classification Results

Class	Precision	Recall	F1-score	Support
Non-Toxic	0.9077	0.9558	0.9311	679
Toxic	0.3617	0.2048	0.2615	83
Accuracy		0.8740		
Macro avg	0.6347	0.5803	0.5963	762
Weighted avg	0.8482	0.8740	0.8582	762

Table: Performance of BERT toxic phrase classification (test set)

High accuracy but low recall on toxic class

Most Toxic Sentence Detection: Confusion Matrix

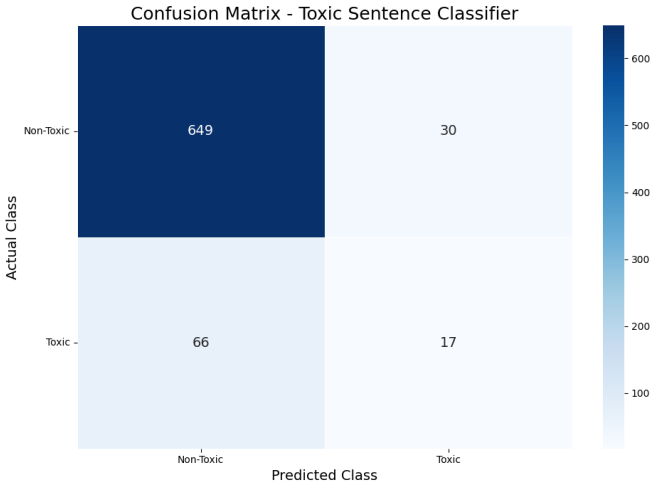

Figure: Confusion matrix - BERT classification model

Most Toxic Sentence Generation: Results

Metric	Score
BLEU	0.1706
ROUGE-1 F1	0.3031
ROUGE-2 F1	0.2170
ROUGE-L F1	0.2987

Table: BART toxic phrase generation performance

Captured some key n-grams, but generation remains challenging

Most Toxic Sentence Generation: Loss Curves

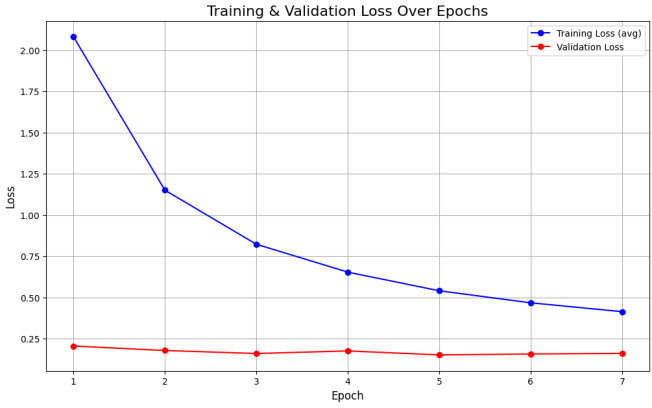

Figure: Training and validation loss - BART generation model

Stable validation loss; no major overfitting detected

Web Application Interface

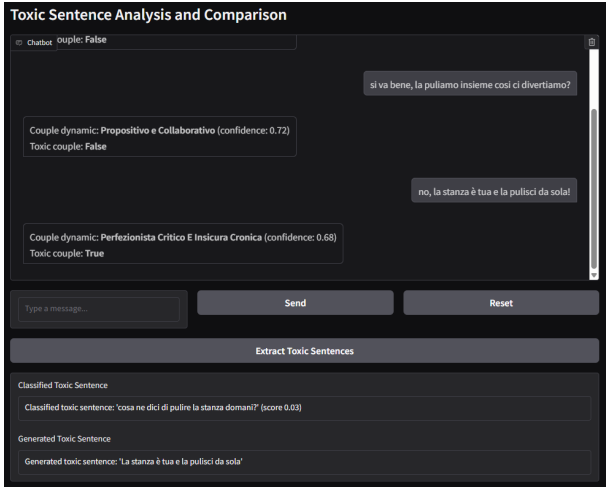

Figure: Screenshot of the interactive web interface

Key Contributions and Results Summary

Key contributions

- Created balanced dataset (~1550 dialogues, annotated toxic phrases).
- Combined toxicity detection and couple dynamics classification for Italian dialogues.
- Integrated classical ML and transformer models for multi-task learning.
- Developed interactive web app for real-time predictions.

Results highlights

- **Binary toxicity classification:** accuracy up to 99.7%.
- **Couple dynamics prediction:** fine-tuned BERT with 80% accuracy.
- **Toxic phrase detection:** high overall accuracy, challenges on minority class.
- **Toxic phrase generation:** achieved moderate results, capturing key n-grams and contextual nuances.

Current Limitations & Future Directions

Current Limitations

- Class imbalance impacted detection performance.
- Transformer models remain hard to interpret.
- Real-time web app may slow down on long inputs.

Future Directions

- Address imbalance: oversampling, focal loss.
- Improve interpretability: attention maps, attribution.
- Optimize inference: distillation, quantization.
- Expand dataset diversity.
- Add explanation features to the web app.

Availability

Code and dataset available on GitHub:
`https://github.com/Davy592/NLP`

Thank You for Your Attention 🚀

Davide Cirilli
d.cirilli2@studenti.uniba.it

Università degli Studi di Bari Aldo Moro