Beyond Labels: Fine-Grained Detection and Explainable AI for Toxicity in Intimate Dialogues

Nicolò Resta

University of Bari, Aldo Moro

EVALITA 2025

Outline

1. Introduction & Motivation
2. Dataset Preparation
3. Experiments
4. Results
5. Conclusion & Future Work

The Challenge: Toxicity in Private Conversations

The Problem

Detecting toxicity (harassment, abuse) is a critical NLP task for online safety. However, most research focuses on **public platforms** (e.g., Twitter, Reddit, Facebook, Instagram, Youtube).

Key Gaps in Private Dialogues

- **Data Scarcity:** Sensitive, private nature makes data acquisition nearly impossible.
- **Context is Paramount:** Toxicity is subtle, dyadic, and depends on relational history.
- **Explainability is Crucial:** Simply flagging a chat as "toxic" is insufficient for trust and intervention. We need to answer questions like *how much?* and *why?*

Our Contributions

1. **A Novel Synthetic Data Generation Pipeline**
   We created a psychologically-grounded pipeline to generate realistic, dyadic **Italian chats**.
   - Annotated with fine-grained, *message-level continuous toxicity scores* [-1, 1].
   - Includes *human-readable narrative explanations* for chat dynamics.
2. **A Comprehensive Empirical Study**
   We conducted experiments across three distinct tasks:
   - Chat-Level Classification (Binary & Multiclass)
   - Message-Level Regression
   - Abstractive Explanation Generation
3. **A Rigorous Comparative Analysis**
   We statistically compared traditional ML models vs. transformer architectures, revealing key insights about the task's inherent limitations.

Our Contributions

1. **A Novel Synthetic Data Generation Pipeline**
   We created a psychologically-grounded pipeline to generate realistic, dyadic **Italian chats**.
   - Annotated with fine-grained, *message-level continuous toxicity scores* [-1, 1].
   - Includes *human-readable narrative explanations* for chat dynamics.
2. **A Comprehensive Empirical Study**
   We conducted experiments across three distinct tasks:
   - Chat-Level Classification (Binary & Multiclass)
   - Message-Level Regression
   - Abstractive Explanation Generation
3. **A Rigorous Comparative Analysis**
   We statistically compared traditional ML models vs. transformer architectures, revealing key insights about the task's inherent limitations.

Our Contributions

1. **A Novel Synthetic Data Generation Pipeline**
   We created a psychologically-grounded pipeline to generate realistic, dyadic **Italian chats**.
   - Annotated with fine-grained, *message-level continuous toxicity scores* [-1, 1].
   - Includes *human-readable narrative explanations* for chat dynamics.
2. **A Comprehensive Empirical Study**
   We conducted experiments across three distinct tasks:
   - Chat-Level Classification (Binary & Multiclass)
   - Message-Level Regression
   - Abstractive Explanation Generation
3. **A Rigorous Comparative Analysis**
   We statistically compared traditional ML models vs. transformer architectures, revealing key insights about the task's inherent limitations.

A 3-Phase Synthetic Data Pipeline

Phase 1, Personas Generation:

- LLM acts as a psychologist.
- Creates detailed profiles: Big Five traits, Attachment styles, Emotional triggers, Personal history, And so on ...

Phase 2, Chat Generation:

- Simulates chats across 7 relationship stages (e.g., infatuation, crisis).
- Controlled toxicity by targeting specific **mean** and **std. dev.** for message scores.
- Each message annotated with a polarity score in $[-1, 1]$.

Phase 3, Explanation Generation:

- LLM acts as a communication expert.
- Generates a narrative rationale explaining the toxic or healthy dynamics of the entire chat.

Result

A rich, nuanced dataset grounded in psychological plausibility, tailored for fine-grained analysis and explainability.

A 3-Phase Synthetic Data Pipeline

Phase 1, Personas Generation:

- LLM acts as a psychologist.
- Creates detailed profiles: Big Five traits, Attachment styles, Emotional triggers, Personal history, And so on ...

Phase 2, Chat Generation:

- Simulates chats across 7 relationship stages (e.g., infatuation, crisis).
- Controlled toxicity by targeting specific **mean** and **std. dev.** for message scores.
- Each message annotated with a polarity score in $[-1, 1]$.

Phase 3, Explanation Generation:

- LLM acts as a communication expert.
- Generates a narrative rationale explaining the toxic or healthy dynamics of the entire chat.

Result

A rich, nuanced dataset grounded in psychological plausibility, tailored for fine-grained analysis and explainability.

A 3-Phase Synthetic Data Pipeline

Phase 1, Personas Generation:

- LLM acts as a psychologist.
- Creates detailed profiles: Big Five traits, Attachment styles, Emotional triggers, Personal history, And so on ...

Phase 2, Chat Generation:

- Simulates chats across 7 relationship stages (e.g., infatuation, crisis).
- Controlled toxicity by targeting specific **mean** and **std. dev.** for message scores.
- Each message annotated with a polarity score in $[-1, 1]$.

Phase 3, Explanation Generation:

- LLM acts as a communication expert.
- Generates a narrative rationale explaining the toxic or healthy dynamics of the entire chat.

Result

A rich, nuanced dataset grounded in psychological plausibility, tailored for fine-grained analysis and explainability.

A 3-Phase Synthetic Data Pipeline

Phase 1, Personas Generation:

- LLM acts as a psychologist.
- Creates detailed profiles: Big Five traits, Attachment styles, Emotional triggers, Personal history, And so on ...

Phase 2, Chat Generation:

- Simulates chats across 7 relationship stages (e.g., infatuation, crisis).
- Controlled toxicity by targeting specific **mean** and **std. dev.** for message scores.
- Each message annotated with a polarity score in $[-1, 1]$.

Phase 3, Explanation Generation:

- LLM acts as a communication expert.
- Generates a narrative rationale explaining the toxic or healthy dynamics of the entire chat.

Result

A rich, nuanced dataset grounded in psychological plausibility, tailored for fine-grained analysis and explainability.

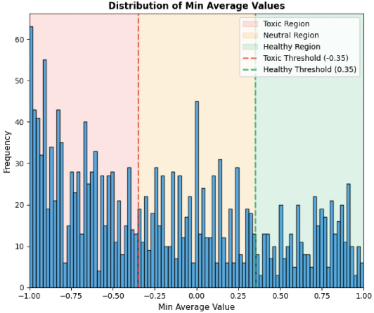
Final Dataset

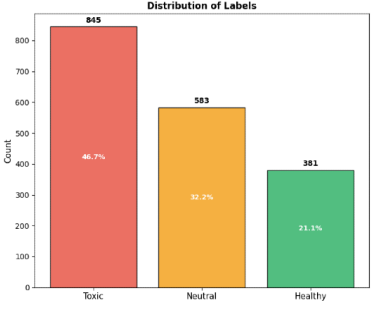
After regex-parsing, filtering (e.g., max length 512 tokens) and cleaning (e.g. converting emojis to text), we obtained **1809** conversations.

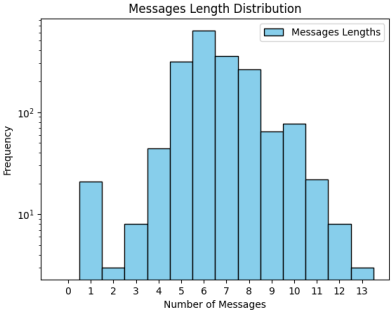
Labeling Scheme for Classification

- **Continuous Message Scores → Discrete Chat Labels**.
- **Multiclass:**
  - Toxic: $[-1, -0.35)$
  - Neutral: $[-0.35, 0.35]$
  - Healthy: $(0.35, 1]$
- **Binary:** Toxic vs. Non-Toxic (Neutral + Healthy).
- **Aggregation:** Chat label is determined by the class of the *minimum* user-average score.

Dataset Distribution







1. Chat-Level Classification

Classic Models:

- Multinomial Naive Bayes (NB)
- Logistic Regression (LR)
- Support Vector Classifier (SVC)

Transformer:

- `dbmdz/bert-base-italian-cased`

2. Message-Level Regression

Transformer:

- BERT to predict continuous message polarity scores.

3. Explanation Generation

Transformer:

- `morenolq/bart-it` to generate narrative explanations.

Classic Models (LR, SVC, NB)

- NLP Pipeline: Tokenization, POS filtering, NER-based anonymization.
- Normalization: Stemming vs. Lemmatization.
- Vectorization: CountVectorizer (NB) and TfidfVectorizer (LR, SVC).

BERT Models We explored various input strategies:

- **BERT (Classification):** Simple concatenation of all messages.
- **BERT-ST (Classification):** Separating speaker turns with '[SEP]' and using token-type IDs.
- **BERT-M (Regression):** Full chat as context, target message wrapped in '[SEP]', distinguished with token-type IDs.
- **BERT-MU (Regression):** Role-aware variant of BERT-M with additional learned user embeddings.

Rigorous Evaluation Protocol

Nested Cross-Validation

To get an unbiased performance estimate and tune hyperparameters simultaneously.

- **Outer Loop (5-fold):** For robust performance estimation.
- **Inner Loop (3-fold):** For hyperparameter tuning (`GridSearchCV`).

Preventing Data Leakage

`GroupKFold` was used in both loops, ensuring all chats from the same couple remain in the same fold. This is crucial for valid results.

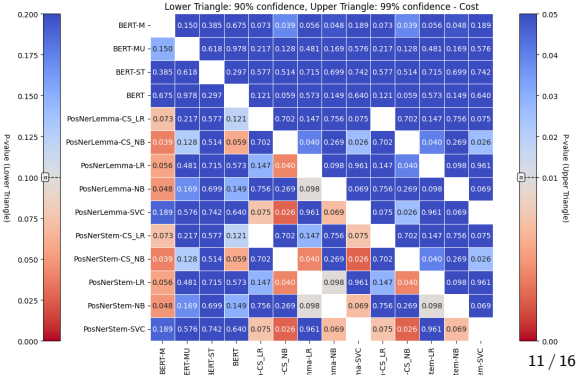
Statistical Analysis

We computed means, stds, 95% confidence intervals, and conducted **paired t-tests** on the outer fold scores to determine if performance differences were statistically significant.

Multiclass Classification (toxic, neutral, healthy)

- Top models converge around a **Weighted F1-score of 0.78**.
- Key Finding: Classic models like **Logistic Regression** and **SVC** perform on par with more complex **BERT** architectures.

Model	Weighted F1	Cost
PosNerStem-LR	0.76 ± 0.02	0.11 ± 0.01
PosNerStem-SVC	0.77 ± 0.03	0.11 ± 0.01
BERT	0.77 ± 0.03	0.10 ± 0.02
BERT-M (Reg.)	$\mathbf{0.78 \pm 0.04}$	$\mathbf{0.10 \pm 0.02}$



Finding 2: Regression is a Viable Alternative

Message-Level Regression Performance

Both BERT regression models accurately predict continuous toxicity scores.

- **Correlation Coefficient ¿ 0.82** with ground truth.
- Substantial improvement over a naive baseline (e.g., R-MAE 0.46).
- The simpler **BERT-M** model slightly outperformed the role-aware **BERT-MU**.

Table: Message-Level Regression Metrics

Metric	BERT-M	BERT-MU
Correlation Coefficient	$\mathbf{0.8335 \pm 0.0264}$	0.8254 ± 0.0266
Relative MAE (R-MAE)	$\mathbf{0.4618 \pm 0.0372}$	0.4687 ± 0.0390

Connecting Regression to Classification

When aggregating regression scores to produce chat-level labels, the **BERT-M** model achieves top-tier classification performance (**0.78 F1** multiclass, **0.87 F1** binary), proving the validity

Finding 3: Promising Results in Explainability

Abstractive Explanation Generation

The BART model was trained to generate narrative summaries of chat dynamics.

Table: Explanation Generation Test Metrics

Metric	Value
BERTScore (F1)	**0.77**
ROUGE-1	0.56
ROUGE-2	0.21
ROUGE-L	0.25
BLEU	0.20

Interpretation

- **High BERTScore F1 (0.77)** indicates strong *semantic similarity* between generated and reference explanations. The model captures the correct meaning.
- Lower n-gram scores (ROUGE-2, BLEU) are expected in abstractive tasks with high linguistic variability.
- Overall: The model can generate coherent and contextually relevant rationales, a crucial step towards trustworthy AI.

Conclusions

- We introduced a novel, psychologically-grounded pipeline for generating rich synthetic data for toxicity analysis in intimate dialogues.
- Our key finding is a **performance plateau**: a diverse range of models (from LR to BERT) achieve statistically similar peak F1-scores (0.78 multiclass, 0.87 binary).
- This suggests performance is currently bottlenecked by the **task's inherent ambiguity** and data characteristics, rather than model complexity.
- The fine-grained regression approach is highly effective and performs competitively on classification tasks when its outputs are aggregated.
- Our BART model demonstrates promising capabilities for generating **semantically relevant explanations**, paving the way for more transparent systems.

Conclusions

- We introduced a novel, psychologically-grounded pipeline for generating rich synthetic data for toxicity analysis in intimate dialogues.
- Our key finding is a **performance plateau**: a diverse range of models (from LR to BERT) achieve statistically similar peak F1-scores (0.78 multiclass, 0.87 binary).
- This suggests performance is currently bottlenecked by the **task's inherent ambiguity** and data characteristics, rather than model complexity.
- The fine-grained regression approach is highly effective and performs competitively on classification tasks when its outputs are aggregated.
- Our BART model demonstrates promising capabilities for generating **semantically relevant explanations**, paving the way for more transparent systems.

Conclusions

- We introduced a novel, psychologically-grounded pipeline for generating rich synthetic data for toxicity analysis in intimate dialogues.
- Our key finding is a **performance plateau**: a diverse range of models (from LR to BERT) achieve statistically similar peak F1-scores (0.78 multiclass, 0.87 binary).
- This suggests performance is currently bottlenecked by the **task's inherent ambiguity** and data characteristics, rather than model complexity.
- The fine-grained regression approach is highly effective and performs competitively on classification tasks when its outputs are aggregated.
- Our BART model demonstrates promising capabilities for generating **semantically relevant explanations**, paving the way for more transparent systems.

Conclusions

- We introduced a novel, psychologically-grounded pipeline for generating rich synthetic data for toxicity analysis in intimate dialogues.
- Our key finding is a **performance plateau**: a diverse range of models (from LR to BERT) achieve statistically similar peak F1-scores (0.78 multiclass, 0.87 binary).
- This suggests performance is currently bottlenecked by the **task's inherent ambiguity** and data characteristics, rather than model complexity.
- The fine-grained regression approach is highly effective and performs competitively on classification tasks when its outputs are aggregated.
- Our BART model demonstrates promising capabilities for generating **semantically relevant explanations**, paving the way for more transparent systems.

Conclusions

- We introduced a novel, psychologically-grounded pipeline for generating rich synthetic data for toxicity analysis in intimate dialogues.
- Our key finding is a **performance plateau**: a diverse range of models (from LR to BERT) achieve statistically similar peak F1-scores (0.78 multiclass, 0.87 binary).
- This suggests performance is currently bottlenecked by the **task's inherent ambiguity** and data characteristics, rather than model complexity.
- The fine-grained regression approach is highly effective and performs competitively on classification tasks when its outputs are aggregated.
- Our BART model demonstrates promising capabilities for generating **semantically relevant explanations**, paving the way for more transparent systems.

Future Work

Our ongoing research focuses on four key areas:

1. **Enhancing Data Generation & Quality Assessment**
   Refining the pipeline with automated quality metrics and exploring multi-agent (e.g., critic/generator) frameworks to improve realism.
2. **Real-World Validation**
   Deploying a public demo to collect user feedback, bridging the "sim-to-real" gap and testing model generalization.
3. **Multi-Task Learning for Enhanced Explainability**
   Training a single BART-based model for both regression and explanation generation, using explanation as a form of regularization to learn more robust representations.
4. **Interdisciplinary Collaboration**
   Integrating professional psychologists into the research team to improve psychological fidelity and validate model behaviors.

Thank You!

Questions?