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

Nicolò Resta

University of Bari, Aldo Moro

September 17, 2025

Outline

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.

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 (OCEAN), Attachment styles, Emotional triggers, Personal history, Motivations, Defensive Mechanisms and so on ...

Phase 2, Chat Generation:

- Simulates chats across 7 relationship stages (e.g., initiation, exploration ...).
- 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 (OCEAN), Attachment styles, Emotional triggers, Personal history, Motivations, Defensive Mechanisms and so on ...

Phase 2, Chat Generation:

- Simulates chats across 7 relationship stages (e.g., initiation, exploration ...).
- 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 (OCEAN), Attachment styles, Emotional triggers, Personal history, Motivations, Defensive Mechanisms and so on ...

Phase 2, Chat Generation:

- Simulates chats across 7 relationship stages (e.g., initiation, exploration ...).
- 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 (OCEAN), Attachment styles, Emotional triggers, Personal history, Motivations, Defensive Mechanisms and so on ...

Phase 2, Chat Generation:

- Simulates chats across 7 relationship stages (e.g., initiation, exploration ...).
- 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.

Preparing the Final Dataset for Experiments

Filtering and Cleaning

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

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

Dataset Structures

- **Classification**: Each sample consists of a textual chat + chat-level label (binary/multiclass).
- **Regression**: Each sample consists of a target message, its message-level score and the full chat as context.

Preparing the Final Dataset for Experiments

Filtering and Cleaning

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

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

Dataset Structures

- **Classification**: Each sample consists of a textual chat + chat-level label (binary/multiclass).
- **Regression**: Each sample consists of a target message, its message-level score and the full chat as context.

Preparing the Final Dataset for Experiments

Filtering and Cleaning

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

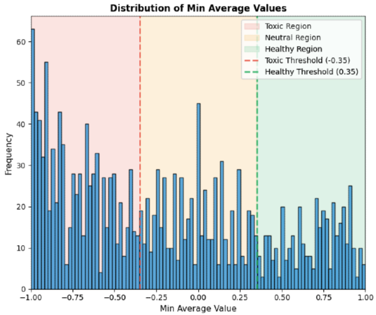
Labeling Scheme for Classification

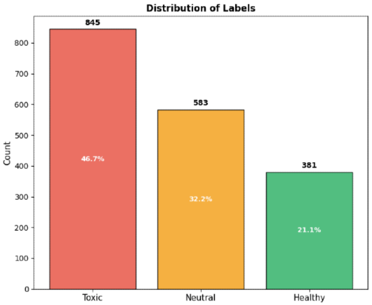
- **Aggregation:** Chat label is determined by the class of the *minimum* user-average score.
- **Multiclass:**
  - Toxic: $[-1, -0.35)$
  - Neutral: $[-0.35, 0.35]$
  - Healthy: $(0.35, 1]$
- **Binary:** Toxic vs. Non-Toxic (Neutral + Healthy).

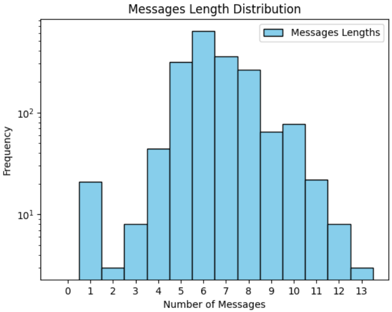
Dataset Structures

- **Classification**: Each sample consists of a textual chat + chat-level label (binary/multiclass).
- **Regression**: Each sample consists of a target message, its message-level score and the full chat as context.

Dataset Distribution







All Models Used

Classic Machine Learning Models

For classification task:

- Tokenization
- POS filtering: Keep only NOUNs, VERBs, ADJs, ADVs, PRONs, AUXs and INTJs.
- NER-based anonymization: Replace PERSON, ORG and LOC.
- Normalization: Stemming vs. Lemmatization.
- Vectorization + Models: CountVectorizer (NB) and TfidfVectorizer (LR, SVC).

Transformer Models

- BERT: `dbmdz/bert-base-italian-cased` for both classification and regression tasks
- BART: `morenolq/bart-it` for explanation task

All Models Used

Classic Machine Learning Models

For classification task:

- Tokenization
- POS filtering: Keep only NOUNs, VERBs, ADJs, ADVs, PRONs, AUXs and INTJs.
- NER-based anonymization: Replace PERSON, ORG and LOC.
- Normalization: Stemming vs. Lemmatization.
- Vectorization + Models: CountVectorizer (NB) and TfidfVectorizer (LR, SVC).

Transformer Models

- BERT: `dbmdz/bert-base-italian-cased` for both classification and regression tasks
- BART: `morenolq/bart-it` for explanation task

Cost-Sensitive Prediction Model Variants

- Logistic Regression and Naive Bayes were also wrapped in a `CostSensitiveClassifier` to make lowest-cost predictions.
- These models were trained and evaluated as standalone models for comparisons.
- Cost matrix was defined based on the psychological severity of misclassifications:

$$M = \begin{bmatrix} 0 & 8 & 16 \\ 8 & 0 & 1 \\ 16 & 4 & 0 \end{bmatrix}$$

BERT Input Formatting Variants for Chat-Level Classification

BERT Input (Simple Concatenation):

[CLS]	User	A:	msg1	User	B:	msg2	User	A:	msg3

BERT-ST Input (Structured with Speaker Information):

[CLS]	User	A:	msg1	[SEP]	User	B:	msg2	[SEP]	User	A:	msg3	[SEP]

Token-type IDs:

0	0	0	0	0	1	1	1	1	0	0	0	0

Token-Type IDs

- Token-type IDs distinguish messages from different users.
- Each user's messages are separated by special tokens ([SEP]).

BERT-M Input (Target Message Highlighted):

[CLS]	User	A:	msg1	[SEP]	User	B:	msg2	[SEP]	User	A:	msg3	User	B:	msg4

Token-type IDs:

1	0	0	0	1	1	1	1	1	0	0	0	0	0	0

Token-Type IDs

- Token-type IDs distinguish target and contextual tokens.
- The target message is isolated between two special tokens ([SEP]).

BERT Input Formatting Variants for Message-Level Regression 2/2

BERT-MU Input (With Learned User Embeddings):

[CLS]	User	A:	msg1	[SEP]	User	B:	msg2	[SEP]	User	A:	msg3	User	B:	msg4

Token-type IDs:

1	0	0	0	1	1	1	1	1	0	0	0	0	0	0

User-type IDs:

0	1	1	1	0	0	0	0	0	1	1	1	0	0	0

User-Type IDs

- Each token also receives a learned user embedding.
- This allows the model to capture user-specific communication styles and patterns.

Rigorous Evaluation Protocol for Classification/Regression

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`) (only for ML models).

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

Evaluation Metrics Across All Tasks

Classification

- **Accuracy**: Ratio of correct predictions
- **Precision**: Accuracy of positive predictions
- **Recall**: Ability to find all positives
- **F1-Score**: Harmonic mean of precision/recall
- **Misclassification Cost**: Custom penalty based on psychological severity (multiclass only)

Regression

- **MAE**: Average absolute error, outlier-robust
- **RMSE - MSE**: Root mean squared error, penalizes large errors
- **Correlation**: Linear relationship strength
- **R-MAE/R-RMSE/R-MSE**: Relative to naive baseline

Explanation Generation

- **ROUGE-1/2**: N-gram overlap (unigrams/bigrams)
- **ROUGE-L**: Longest common subsequence
- **BLEU**: N-gram precision-focused
- **BERTScore**: Semantic similarity using BERT embeddings. Provides P/R/F1 for robust semantic evaluation

Hyperparameters Space

Component	Hyperparameter	Values
Count/Tfidf Vectorizer	ngram_range	(1, 1), (1, 2), (1, 3)
	min_df	3, 8, 20
	max_df	0.9, 0.95, 0.99
Multinomial NB	alpha	0.1, 0.5, 1.0, 2.0
Logistic Regression	C	0.1, 1.0, 10.0
	max_iter	1000, 2000
BERT	n. Max. Epochs	20
	Learning Rate	3e-5
	Batch Size	32
	Grad. Accum. Steps	4
	Weight Decay	0.001
	Warmup Percentage	0.1
	Early Stopping	Patience: 4
	LR Scheduler	Reduce on Plateau
	(Factor: 0.5,	Patience: 2)

Component	Hyperparameter	Values
BART	n. Max. Epochs	20
	Learning Rate	3e-5
	Batch Size	4
	Grad. Accum. Steps	8
	Weight Decay	0.01
	Warmup Percentage	0.1
	LR Scheduler	Linear with Warmup

Best Model Selection Criteria

Chat-Level Classification Tasks

- Binary: Maximize **Weighted F1-score**.
- Multiclass: Minimize **Chat-Level Misclassification Cost** of chat-level predictions.

Message-Level Regression Task

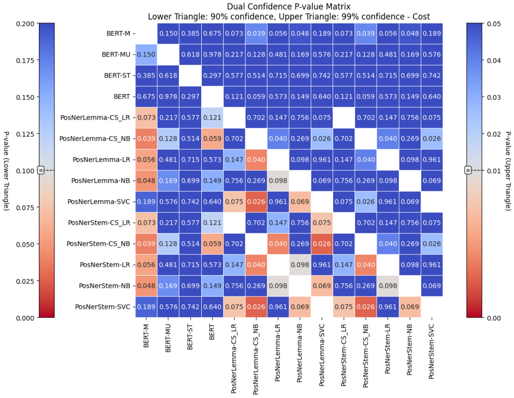
Minimize **Message-Level Misclassification Cost** of aggregated message-level predictions.

Explanation Generation Task

Maximize **BERTScore F1** between generated and reference explanations.

Finding 1: A Performance Plateau in Classification (Multiclass)

Model	Weighted F1	Cost
BERT-M	$0.78 \pm 0.04\ [0.72, 0.83]$	$0.10 \pm 0.02\ [0.07, 0.12]$
BERT	$0.77 \pm 0.03\ [0.72, 0.81]$	$0.10 \pm 0.02\ [0.07, 0.13]$
BERT-MU	$0.77 \pm 0.05\ [0.71, 0.84]$	$0.10 \pm 0.02\ [0.07, 0.13]$
PosNerStem-SVC	$0.77 \pm 0.03\ [0.73, 0.80]$	$0.11 \pm 0.01\ [0.09, 0.13]$
PosNerLemma-SVC	$0.77 \pm 0.03\ [0.73, 0.80]$	$0.11 \pm 0.01\ [0.09, 0.13]$
PosNerStem-LR	$0.76 \pm 0.02\ [0.73, 0.79]$	$0.11 \pm 0.01\ [0.09, 0.12]$
PosNerLemma-LR	$0.76 \pm 0.02\ [0.73, 0.79]$	$0.11 \pm 0.01\ [0.09, 0.12]$
BERT-ST	$0.76 \pm 0.04\ [0.70, 0.81]$	$0.11 \pm 0.02\ [0.08, 0.14]$
PosNerStem-CS_NB	$0.75 \pm 0.03\ [0.70, 0.79]$	$0.12 \pm 0.02\ [0.10, 0.15]$
PosNerStem-NB	$0.75 \pm 0.04\ [0.70, 0.80]$	$0.12 \pm 0.02\ [0.09, 0.15]$
PosNerLemma-CS_NB	$0.75 \pm 0.03\ [0.70, 0.79]$	$0.12 \pm 0.02\ [0.10, 0.15]$
PosNerLemma-NB	$0.75 \pm 0.04\ [0.70, 0.80]$	$0.12 \pm 0.02\ [0.09, 0.15]$
PosNerStem-CS_LR	$0.73 \pm 0.03\ [0.68, 0.77]$	$0.12 \pm 0.02\ [0.09, 0.15]$
PosNerLemma-CS_LR	$0.73 \pm 0.03\ [0.68, 0.77]$	$0.12 \pm 0.02\ [0.09, 0.15]$

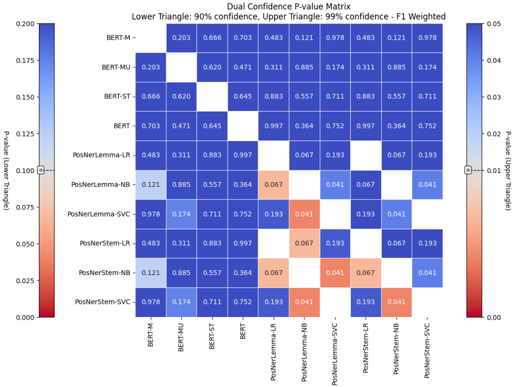


Interpretation

The task's and labeling scheme's inherent **subjectivity and ambiguity** may have challenged all models equally limiting performances.

Finding 1: A Performance Plateau in Classification (Binary)

Model	Weighted F1
SVC	0.87 ± 0.03 [0.83, 0.91]
BERT	0.87 ± 0.02 [0.84, 0.90]
BERT-M	0.87 ± 0.02 [0.84, 0.90]
LR	0.86 ± 0.02 [0.83, 0.90]
BERT-ST	0.86 ± 0.02 [0.82, 0.89]
BERT-MU	0.86 ± 0.03 [0.82, 0.89]
NB	0.84 ± 0.03 [0.81, 0.88]



Interpretation

Mistakenly introduced dataset latent biases may be learnable up to a certain point, beyond which further complex models yields diminishing returns.

Finding 2: Regression is a Viable Alternative

Metric	BERT-M	BERT-MU
Mean Squared Error (MSE)	0.1367 ± 0.0168 [0.1159, 0.1576]	0.1423 ± 0.0194 [0.1182, 0.1664]
Mean Absolute Error (MAE)	0.2656 ± 0.0216 [0.2388, 0.2925]	0.2695 ± 0.0223 [0.2419, 0.2972]
Root Mean Squared Error (RMSE)	0.3692 ± 0.0225 [0.3413, 0.3972]	0.3765 ± 0.0257 [0.3446, 0.4085]
Correlation Coefficient	0.8335 ± 0.0264 [0.8007, 0.8663]	0.8254 ± 0.0266 [0.7923, 0.8584]
Relative MSE (R-MSE)	0.3205 ± 0.0391 [0.2720, 0.3691]	0.3338 ± 0.0467 [0.2758, 0.3918]
Relative MAE (R-MAE)	0.4618 ± 0.0372 [0.4156, 0.5080]	0.4687 ± 0.0390 [0.4203, 0.5171]
Relative RMSE (R-RMSE)	0.5653 ± 0.0342 [0.5229, 0.6078]	0.5766 ± 0.0406 [0.5261, 0.6270]
Message-Level Binary Weighted F1	0.84 ± 0.02 [0.82, 0.86]	0.83 ± 0.02 [0.80, 0.86]
Message-Level Multiclass Weighted F1	0.72 ± 0.03 [0.68, 0.76]	0.71 ± 0.03 [0.67, 0.75]
Message-Level Multiclass Cost	0.13 ± 0.02 [0.10, 0.15]	0.13 ± 0.02 [0.11, 0.15]

Interpretation

Models are effective at capturing the nuances of toxicity in messages. They also achieved competitive classification performances, validating the fine-grained approach.

Finding 3: Promising Results in Explainability

Metric	Value
BERTScore (F1)	**0.77**
BERTScore (Precision)	0.77
BERTScore (Recall)	0.76
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.

Considerations

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 demonstrated promising competitively performances 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 demonstrated promising competitively performances 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 demonstrated promising competitively performances 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 demonstrated promising competitively performances 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 demonstrated promising competitively performances 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. **Interdisciplinary Collaboration**
   Integrating professional psychologists into the research team to improve psychological fidelity and validate model behaviors.
3. **Real-World Validation**
   Deploying a public demo to collect user feedback, bridging the "sim-to-real" gap and testing model generalization.
4. **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.

Thank You!

Nicolò Resta
nicolo.resta@studenti.uniba.it
University of Bari, Aldo Moro