# TOXICITY AND CONVERSATION CATEGORY CLASSIFICATION IN ITALIAN DIALOGUES: AN INTEGRATED NLP APPROACH WITH FINE-TUNED TRANSFORMERS AND WEB APPLICATION

ANGELANTONIO FEDELE MUROLO MAT.840167

UNIVERSITY OF BARI ALDO MORO – COMPUTER SCIENCE ARTIFICAL INTELLIGENCE

NATURAL LANGUAGE PROCESSING - CIPV

# PRESENTATION AGENDA

- Introduction and Motivation
- Tools Used
- Datasets and Preprocessing
- Metodology: Classification and Generative Models
- Evaluation Metrics
- Key Results: Classification and Generation
- PoisonChat: The Web Application
- Conclusions and Limitations
- Future Improvements

# INTRODUCTION AND MOTIVATION

- Online communication platforms face increasing challenges from harmful content, including hate speech and general toxicity.
- The issue is particularly complex for the Italian language due to its linguistic nuances and limited availability of annotated datasets.

- Classify Italian conversations as "toxic" or "non-toxic" and categorize them into granular types, offering deeper insights into dialogue nature.
- Utilize both traditional machine learning methods as Logistic Regression with the TF-IDF and FastText as baselines, and fine-tuned state-of-the-art Transformer models like BERT for classification and BART/T5 for sentence generation/extraction.

# TOOLS USED

Hugging Face

Gemini

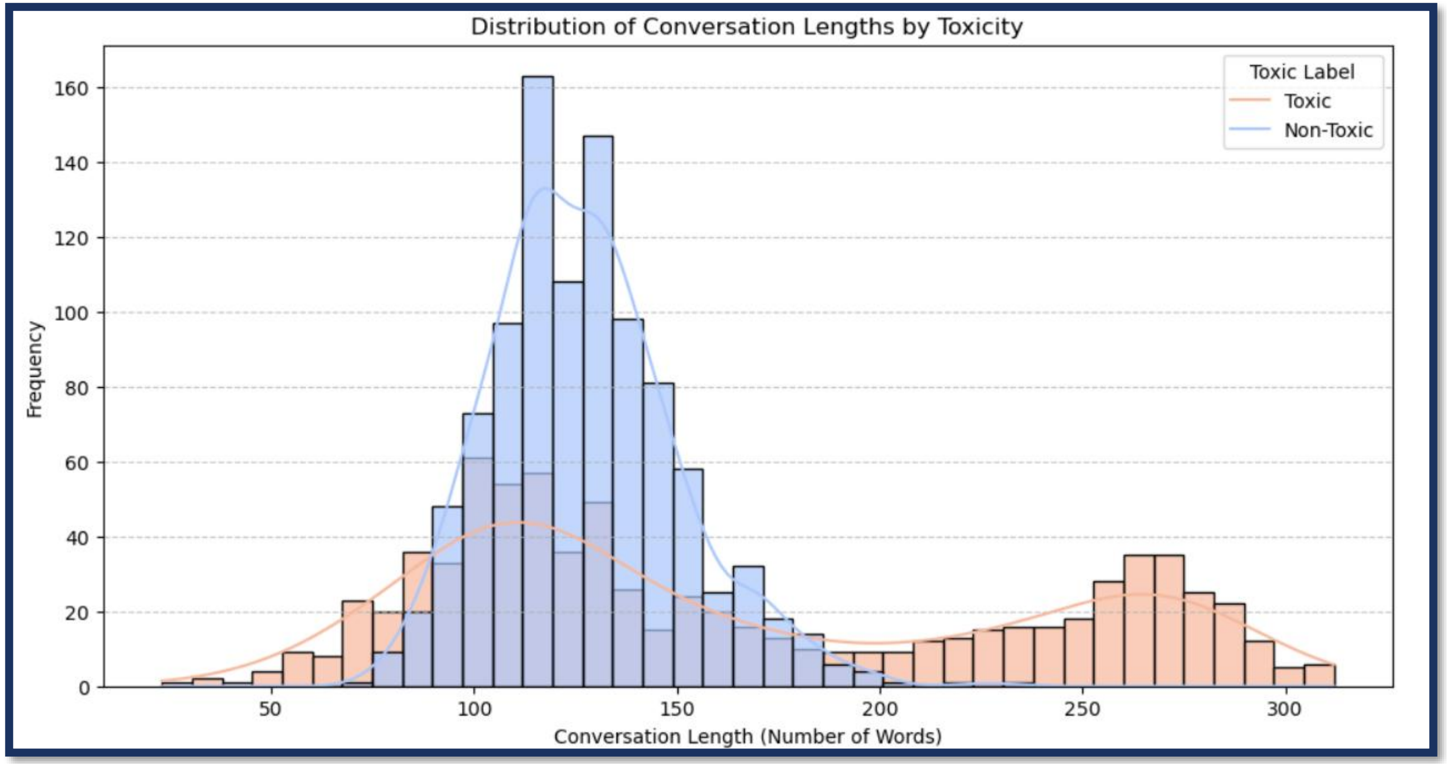
Streamlit

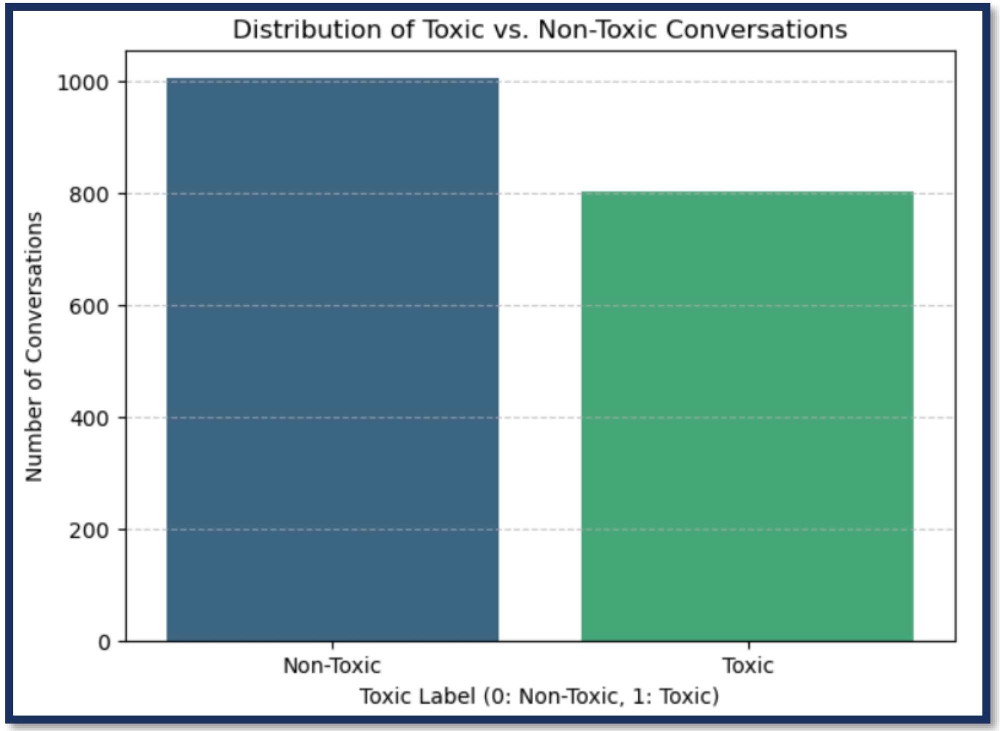
kaggle

# DATASETS AND PREPROCESSING

- **Non-Toxic Data Generation:** Programmatically generated a custom corpus of diverse, healthy Italian dialogues via the Gemini API, ensuring content purity.
- **Toxic Data Integration & Preprocessing:** Merged with an existing toxic Italian dialogue dataset. Specific preprocessing steps included:
  - Removal of incomplete / malformed entries.
  - Extraction of the "most toxic sentence" via regular expressions.
  - Normalization of whitespace, punctuation, and symbols.
- **Dataset Unification & Balancing:** The two datasets were unified and balanced to ensure robust model training.

# DATASETS AND PREPROCESSING





# METHODOLOGY - CLASSIFICATION MODELS

- **Binary Toxicity Classification:**
  - **Traditional ML Models:** Logistic Regression (LR), Support Vector Machines (SVM), Naive Bayes (NB), and Light Gradient Boosting Machine (LightGBM).
  - **Feature Representations:** Evaluated with TF-IDF and FastText embeddings.
- **Multi-class Conversation Categorization:**
  - **Fine-tuned Transformer:** Italian BERT model, specifically fine-tuned for this task.
  - **Embeddings-based Approaches:** BERT's contextual embeddings and RoBERTa's embeddings combined with Logistic Regression.
- **Most Toxic Sentence Classification:**
  - **Fine-tuned Transformer:** The same Italian BERT model used for other classification tasks was fine-tuned for this binary classification (Toxic Sentence vs. Non-Toxic Sentence).

# METHODOLOGY - GENERATIVE MODELS

- **Objective:** To either extract an existing toxic sentence from a dialogue or generate a novel one, representing the most toxic utterance.

- **Models Used:**
  - **T5** (Text-to-Text Transfer Transformer) fine-tuned for this sequence-to-sequence task.
  - **BART** (Bidirectional and Auto-Regressive Transformers) fine-tuned for this sequence-to-sequence task.

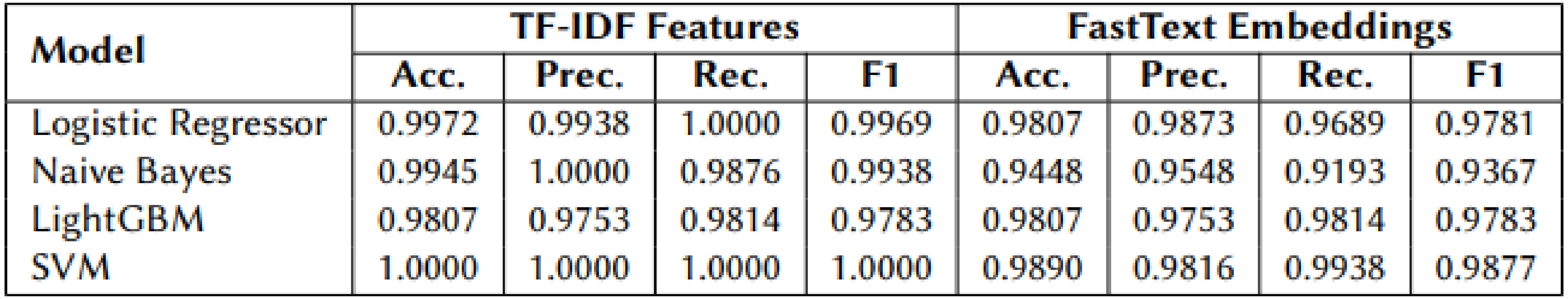
# EVALUATION METRICS

- **For Classification Tasks (Binary Toxicity, Multi-class Categories, Most Toxic Sentence Classification):**
  - **Accuracy:** Overall correctness of predictions.
  - **Precision:** Proportion of true positive predictions among all positive predictions.
  - **Recall:** Proportion of true positive predictions among all actual positives.
  - **F1-score:** Harmonic mean of Precision and Recall, particularly useful for imbalanced datasets.

- **For Generative Task (Most Toxic Sentence Generation):**
  - **BLEU** (Bilingual Evaluation Understudy): Measures the n-gram overlap between generated and reference sentences.
  - **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation): Focuses on recall and overlap of n-grams or sequences, commonly used for summarization and generation tasks (specifically ROUGE-1 and ROUGE-L were used).

# KEY RESULTS - TRADITIONAL ML CLASSIFICATION

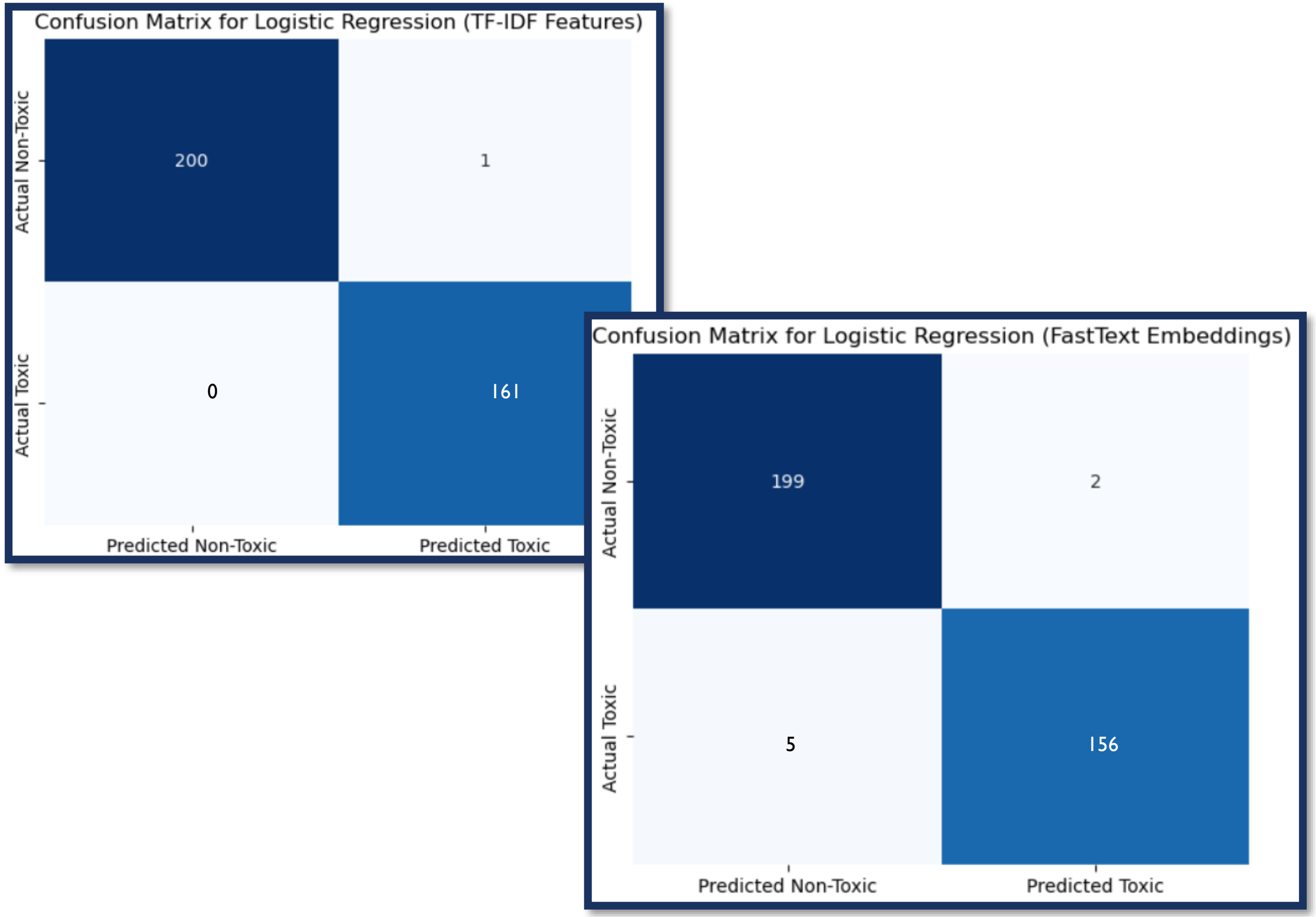
Comparative Performance Metrics for Traditional ML Models (Binary Toxicity Classification)



| Model | TF-IDF Features | | | | FastText Embeddings | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 |
| Logistic Regressor | 0.9972 | 0.9938 | 1.0000 | 0.9969 | 0.9807 | 0.9873 | 0.9689 | 0.9781 |
| Naive Bayes | 0.9945 | 1.0000 | 0.9876 | 0.9938 | 0.9448 | 0.9548 | 0.9193 | 0.9367 |
| LightGBM | 0.9807 | 0.9753 | 0.9814 | 0.9783 | 0.9807 | 0.9753 | 0.9814 | 0.9783 |
| SVM | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 0.9890 | 0.9816 | 0.9938 | 0.9877 |

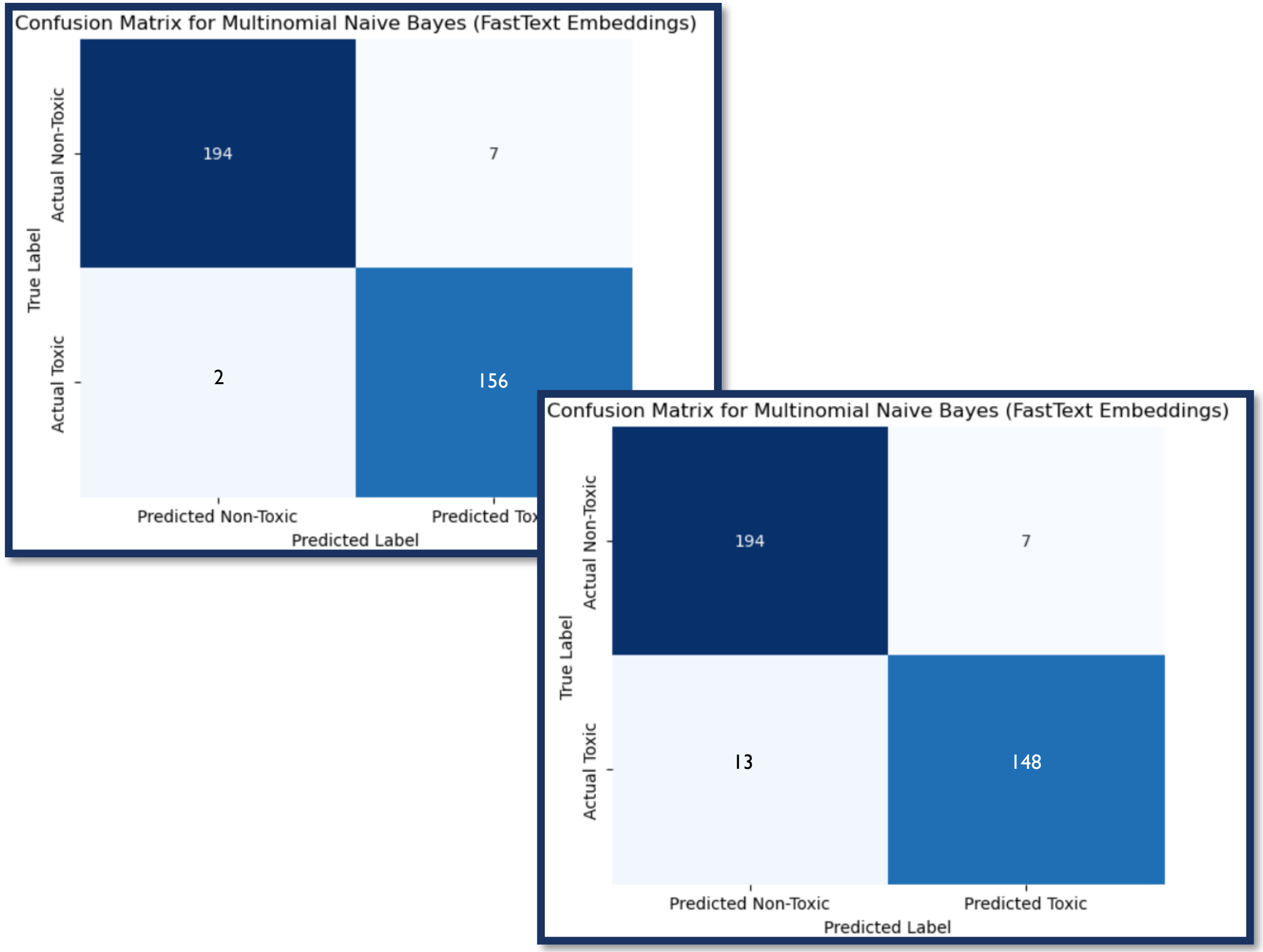
# KEY RESULTS - LOGISTIC REGRESSOR (BINARY TOXICITY)

| Metric | TF-IDF Features | FastText Embeddings |
|---|---|---|
| Accuracy | **0.9972** | 0.9807 |
| Precision | **0.9938** | 0.9873 |
| Recall | **1.0000** | 0.9689 |
| F1-Score | **0.9969** | 0.9781 |



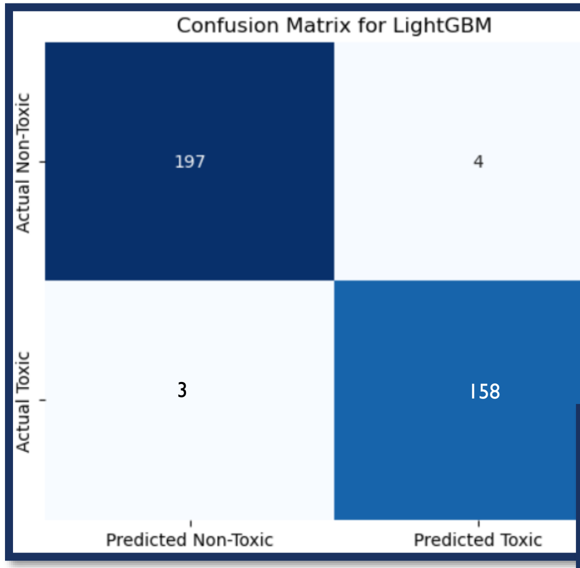
# KEY RESULTS - NAIVE BAYES (BINARY TOXICITY)

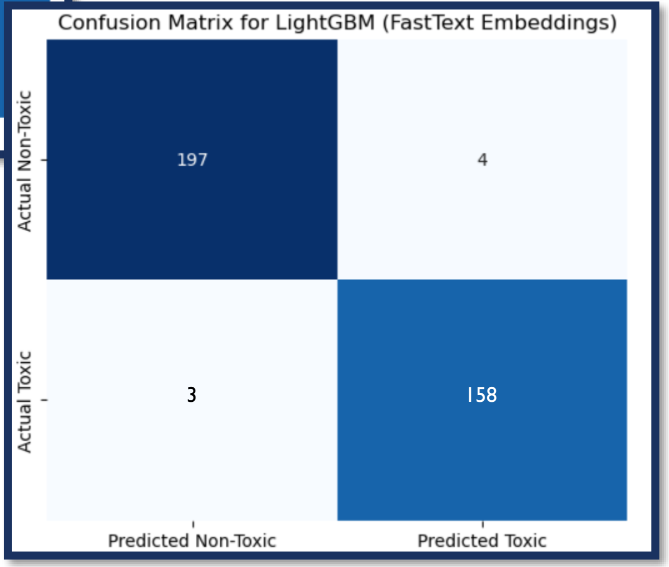
| **Metric** | **TF-IDF Features** | **FastText Embeddings** |
|---|---|---|
| Accuracy | **0.9945** | 0.9448 |
| Precision | **1.0000** | 0.9548 |
| Recall | **0.9876** | 0.9193 |
| F1-Score | **0.9938** | 0.9367 |



# KEY RESULTS - LIGHTGBM (BINARY TOXICITY)

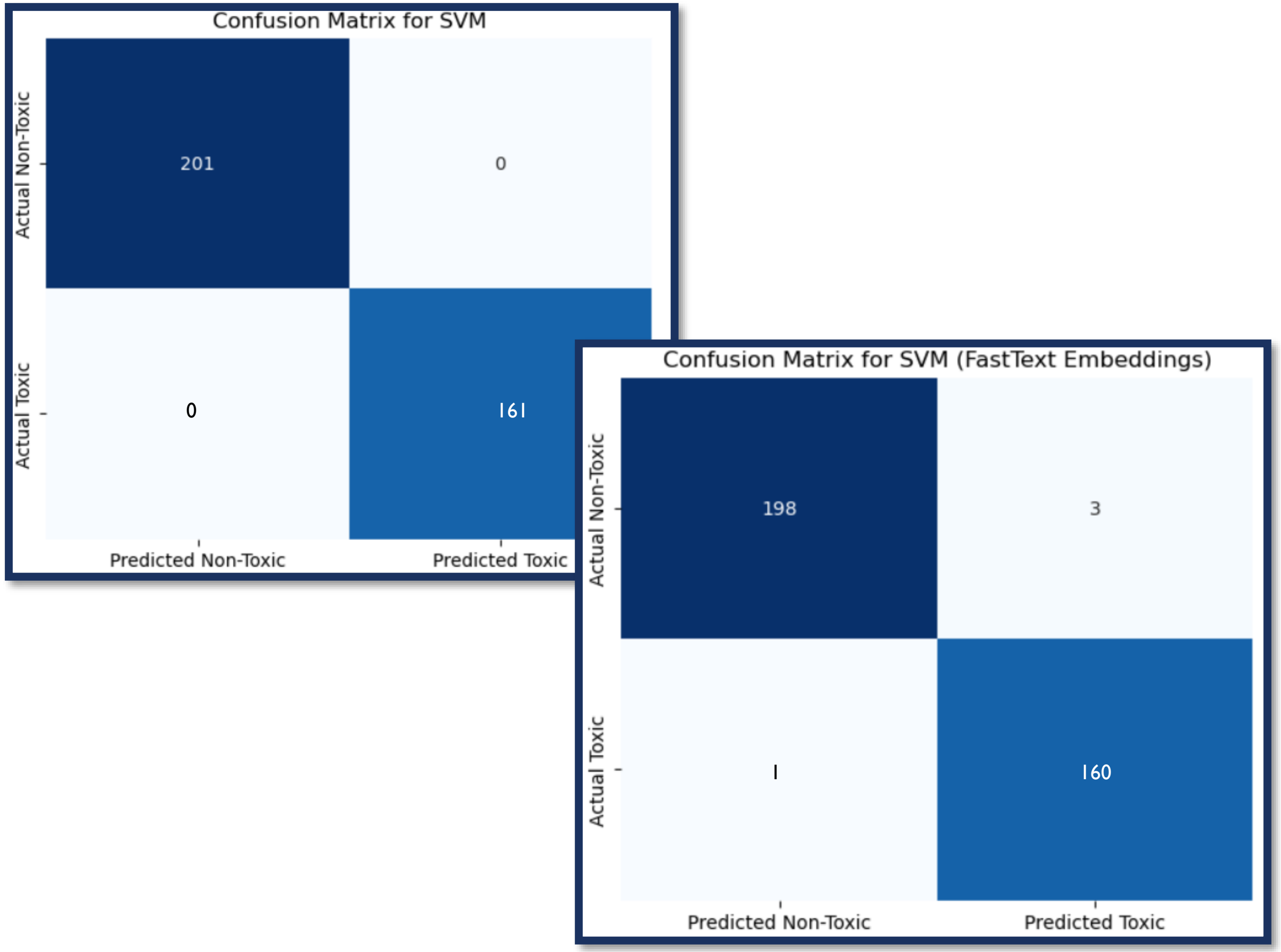
| Metric | TF-IDF Features | FastText Embeddings |
|---|---|---|
| Accuracy | 0.9807 | 0.9807 |
| Precision | 0.9753 | 0.9753 |
| Recall | 0.9814 | 0.9814 |
| F1-Score | 0.9783 | 0.9783 |





# KEY RESULTS - SVM (BINARY TOXICITY)

| Metric | TF-IDF Features | FastText Embeddings |
|---|---|---|
| Accuracy | 1.0000 | 0.9890 |
| Precision | 1.0000 | 0.9816 |
| Recall | 1.0000 | 0.9938 |
| F1-Score | 1.0000 | 0.9877 |



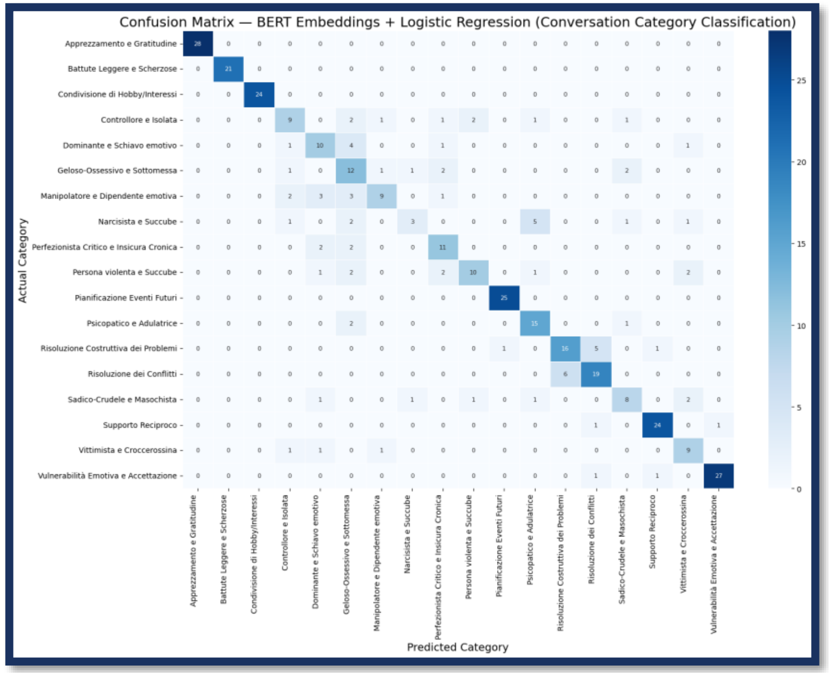
Confusion Matrix for SVM

Confusion Matrix for SVM (FastText Embeddings)

# KEY RESULTS - TRANSFORMER CLASSIFICATION

| **Class** | **BERT (Acc: 0.7735)** | **BERT Fine-tuned (Acc: 0.8232)** | **XLM-RoBERTa (Acc: 0.7643)** |
|---|---|---|---|
| | **P / R / F1** | **P / R / F1** | **P / R / F1** |
| Apprezzamento e Gratitudine | 1.00 / 1.00 / 1.00 | 1.00 / 1.00 / 1.00 | 1.00 / 1.00 / 1.00 |
| Battute Leggere e Scherzose | 1.00 / **1.00** / **1.00** | 1.00 / 0.90 / 0.95 | 1.00 / **1.00** / **1.00** |
| Condivisione di Hobby/Interessi | **1.00** / 1.00 / **1.00** | 0.92 / 1.00 / 0.96 | **1.00** / 1.00 / **1.00** |
| Controllore e Isolata | 0.60 / **0.53** / **0.56** | **1.00** / 0.12 / 0.22 | **1.00** / 0.38 / 0.55 |
| Dominante e Schiavo emotivo | 0.56 / 0.59 / 0.57 | **0.88** / **0.78** / **0.82** | 0.70 / **0.78** / 0.74 |
| Geloso-Ossessivo e Sottomessa | 0.41 / 0.63 / 0.50 | **0.64** / 0.70 / **0.67** | 0.53 / **0.80** / 0.64 |
| Manipolatore e Dipendente emotiva | **0.75** / 0.50 / **0.60** | 0.50 / **0.56** / 0.53 | 0.60 / 0.33 / 0.43 |
| Narcisista e Succube | 0.60 / 0.23 / 0.33 | **1.00** / **0.43** / **0.60** | 0.00 / 0.00 / 0.00 |
| Perfezionista Critico e Insicura Cronica | **0.61** / **0.73** / **0.67** | 0.56 / 0.71 / 0.63 | 0.60 / 0.43 / 0.50 |
| Persona violenta e Succube | **0.77** / 0.56 / **0.65** | 0.55 / **0.67** / 0.60 | 0.42 / 0.56 / 0.48 |
| Pianificazione Eventi Futuri | 0.96 / 1.00 / 0.98 | **1.00** / 1.00 / **1.00** | **1.00** / 1.00 / **1.00** |
| Psicopatico e Adulatrice | 0.65 / 0.83 / 0.73 | **0.82** / **1.00** / **0.90** | 0.47 / 0.89 / 0.62 |
| Risoluzione Costruttiva dei Problemi | 0.73 / 0.70 / 0.71 | **0.91** / **0.91** / **0.91** | 0.90 / 0.82 / 0.86 |
| Risoluzione dei Conflitti | 0.73 / 0.76 / 0.75 | **0.92** / **0.92** / **0.92** | 0.79 / **0.92** / 0.85 |
| Sadico-Crudele e Masochista | **0.62** / **0.57** / **0.59** | 0.44 / **0.57** / 0.50 | 0.00 / 0.00 / 0.00 |
| Supporto Reciproco | 0.92 / 0.92 / 0.92 | **0.93** / **1.00** / **0.96** | **0.93** / **1.00** / **0.96** |
| Vittimista e Croccerossina | 0.60 / 0.75 / 0.67 | **0.75** / **1.00** / **0.86** | 0.36 / 0.67 / 0.47 |
| Vulnerabilità Emotiva e Accettazione | 0.96 / 0.93 / 0.95 | **1.00** / 0.93 / **0.97** | 0.93 / 0.93 / 0.93 |

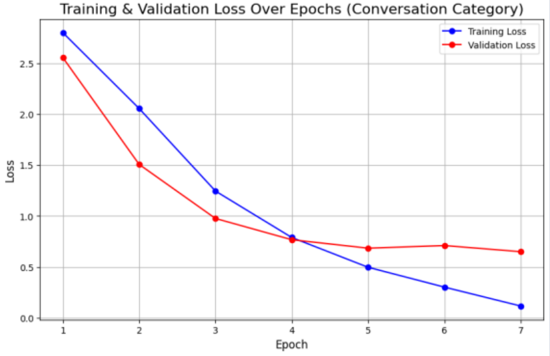
# KEY RESULTS - TRANSFORMER CLASSIFICATION



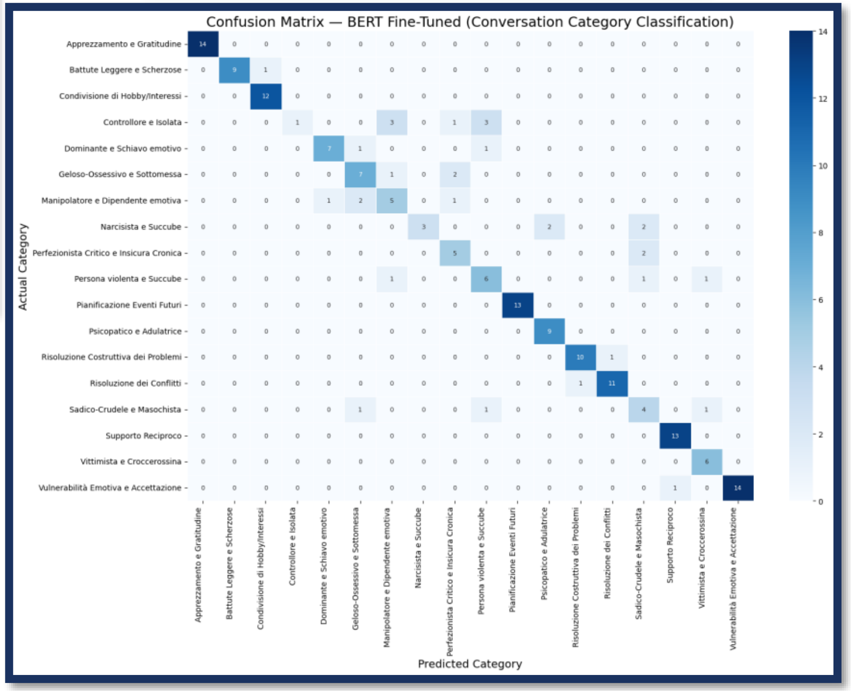
Confusion Matrix — BERT Embeddings + Logistic Regression (Conversation Category Classification)

Actual Category / Predicted Category

| | Apprezzamento e Gratitudine | Battute Leggere e Scherzose | Condivisione di Hobby/Interessi | Controllore e Isolata | Dominante e Schiavo emotivo | Geloso-Ossessivo e Sottomessa | Manipolatore e Dipendente emotiva | Narcisista e Succube | Perfezionista Critico e Insicura Cronica | Persona violenta e Succube | Pianificazione Eventi Futuri | Psicopatico e Adulatrice | Risoluzione Costruttiva dei Problemi | Risoluzione dei Conflitti | Sadico-Crudele e Masochista | Supporto Reciproco | Vittimista e Croccerossina | Vulnerabilità Emotiva e Accettazione |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Apprezzamento e Gratitudine | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Battute Leggere e Scherzose | 0 | 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Condivisione di Hobby/Interessi | 0 | 0 | 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Controllore e Isolata | 0 | 0 | 0 | 9 | 0 | 2 | 1 | 0 | 1 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| Dominante e Schiavo emotivo | 0 | 0 | 0 | 1 | 10 | 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Geloso-Ossessivo e Sottomessa | 0 | 0 | 0 | 1 | 0 | 12 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| Manipolatore e Dipendente emotiva | 0 | 0 | 0 | 2 | 3 | 3 | 9 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Narcisista e Succube | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 3 | 0 | 0 | 0 | 5 | 0 | 0 | 1 | 0 | 1 | 0 |
| Perfezionista Critico e Insicura Cronica | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Persona violenta e Succube | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 2 | 10 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 0 |
| Pianificazione Eventi Futuri | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Psicopatico e Adulatrice | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 1 | 0 | 0 | 0 |
| Risoluzione Costruttiva dei Problemi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 16 | 5 | 0 | 1 | 0 | 0 |
| Risoluzione dei Conflitti | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 19 | 0 | 0 | 0 | 0 |
| Sadico-Crudele e Masochista | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 8 | 0 | 2 | 0 |
| Supporto Reciproco | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 24 | 0 | 1 |
| Vittimista e Croccerossina | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 0 |
| Vulnerabilità Emotiva e Accettazione | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 27 |



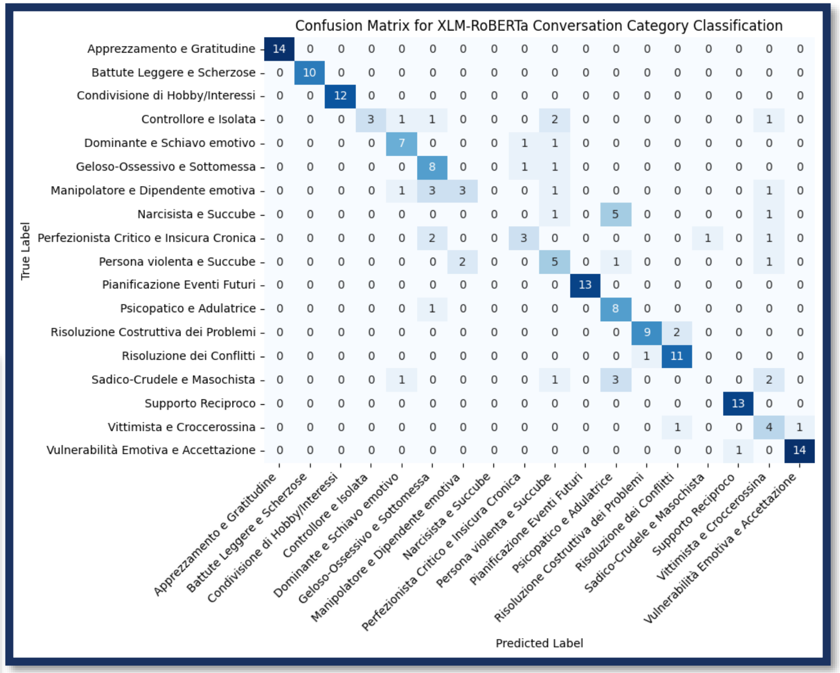
Training & Validation Loss Over Epochs (Conversation Category)



Confusion Matrix — BERT Fine-Tuned (Conversation Category Classification)

Actual Category / Predicted Category

| | Apprezzamento e Gratitudine | Battute Leggere e Scherzose | Condivisione di Hobby/Interessi | Controllore e Isolata | Dominante e Schiavo emotivo | Geloso-Ossessivo e Sottomessa | Manipolatore e Dipendente emotiva | Narcisista e Succube | Perfezionista Critico e Insicura Cronica | Persona violenta e Succube | Pianificazione Eventi Futuri | Psicopatico e Adulatrice | Risoluzione Costruttiva dei Problemi | Risoluzione dei Conflitti | Sadico-Crudele e Masochista | Supporto Reciproco | Vittimista e Croccerossina | Vulnerabilità Emotiva e Accettazione |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Apprezzamento e Gratitudine | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Battute Leggere e Scherzose | 0 | 9 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Condivisione di Hobby/Interessi | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Controllore e Isolata | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Dominante e Schiavo emotivo | 0 | 0 | 0 | 0 | 7 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Geloso-Ossessivo e Sottomessa | 0 | 0 | 0 | 0 | 0 | 7 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Manipolatore e Dipendente emotiva | 0 | 0 | 0 | 0 | 1 | 2 | 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Narcisista e Succube | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 |
| Perfezionista Critico e Insicura Cronica | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| Persona violenta e Succube | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| Pianificazione Eventi Futuri | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Psicopatico e Adulatrice | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| Risoluzione Costruttiva dei Problemi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 1 | 0 | 0 | 0 | 0 |
| Risoluzione dei Conflitti | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 11 | 0 | 0 | 0 | 0 |
| Sadico-Crudele e Masochista | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 4 | 0 | 1 | 0 |
| Supporto Reciproco | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 0 |
| Vittimista e Croccerossina | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |
| Vulnerabilità Emotiva e Accettazione | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 14 |



Confusion Matrix for XLM-RoBERTa Conversation Category Classification

True Label / Predicted Label

| | Apprezzamento e Gratitudine | Battute Leggere e Scherzose | Condivisione di Hobby/Interessi | Controllore e Isolata | Dominante e Schiavo emotivo | Geloso-Ossessivo e Sottomessa | Manipolatore e Dipendente emotiva | Narcisista e Succube | Perfezionista Critico e Insicura Cronica | Persona violenta e Succube | Pianificazione Eventi Futuri | Psicopatico e Adulatrice | Risoluzione Costruttiva dei Problemi | Risoluzione dei Conflitti | Sadico-Crudele e Masochista | Supporto Reciproco | Vittimista e Croccerossina | Vulnerabilità Emotiva e Accettazione |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Apprezzamento e Gratitudine | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Battute Leggere e Scherzose | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Condivisione di Hobby/Interessi | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Controllore e Isolata | 0 | 0 | 0 | 3 | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Dominante e Schiavo emotivo | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Geloso-Ossessivo e Sottomessa | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Manipolatore e Dipendente emotiva | 0 | 0 | 0 | 0 | 1 | 3 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Narcisista e Succube | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 5 | 0 | 0 | 0 | 0 | 1 | 0 |
| Perfezionista Critico e Insicura Cronica | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| Persona violenta e Succube | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| Pianificazione Eventi Futuri | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Psicopatico e Adulatrice | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| Risoluzione Costruttiva dei Problemi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 2 | 0 | 0 | 0 | 0 |
| Risoluzione dei Conflitti | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 11 | 0 | 0 | 0 | 0 |
| Sadico-Crudele e Masochista | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 0 |
| Supporto Reciproco | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 0 | 0 |
| Vittimista e Croccerossina | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 4 | 1 |
| Vulnerabilità Emotiva e Accettazione | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 14 |

# KEY RESULTS - BERT MOST TOXIC SENTENCE CLASSIFICATION

| Class/Average | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Non-Toxic Sentence | 0.92 | 1.00 | 0.96 | 547 |
| Toxic Sentence | 0.89 | 0.95 | 0.92 | 50 |

All sentences and their toxic scores:

'Ciao, come stai?' (Score: 0.0061)

**'Sei un completo idiota e non capisci niente!' (Score: 0.0837)**

'Forse dovremmo parlarne con calma.' (Score: 0.0075)

'Spero tu abbia una buona giornata.' (Score: 0.0099)

**Conversation Category:** 'Litigio’

**Original Conversation:** "Ciao, come stai?", "Sei un completo idiota e non capisci niente!", "Forse dovremmo parlarne con calma.", "Spero tu abbia una buona giornata."

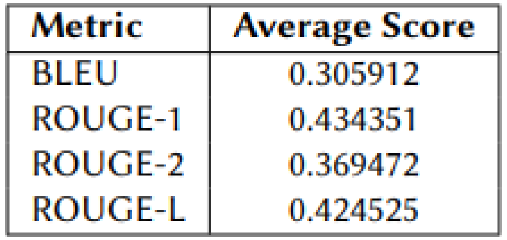
**Identified Most Toxic Sentence:** 'Sei un completo idiota e non capisci niente!' (Toxic Score: 0.0837)

# KEY RESULTS - BART GENERATION

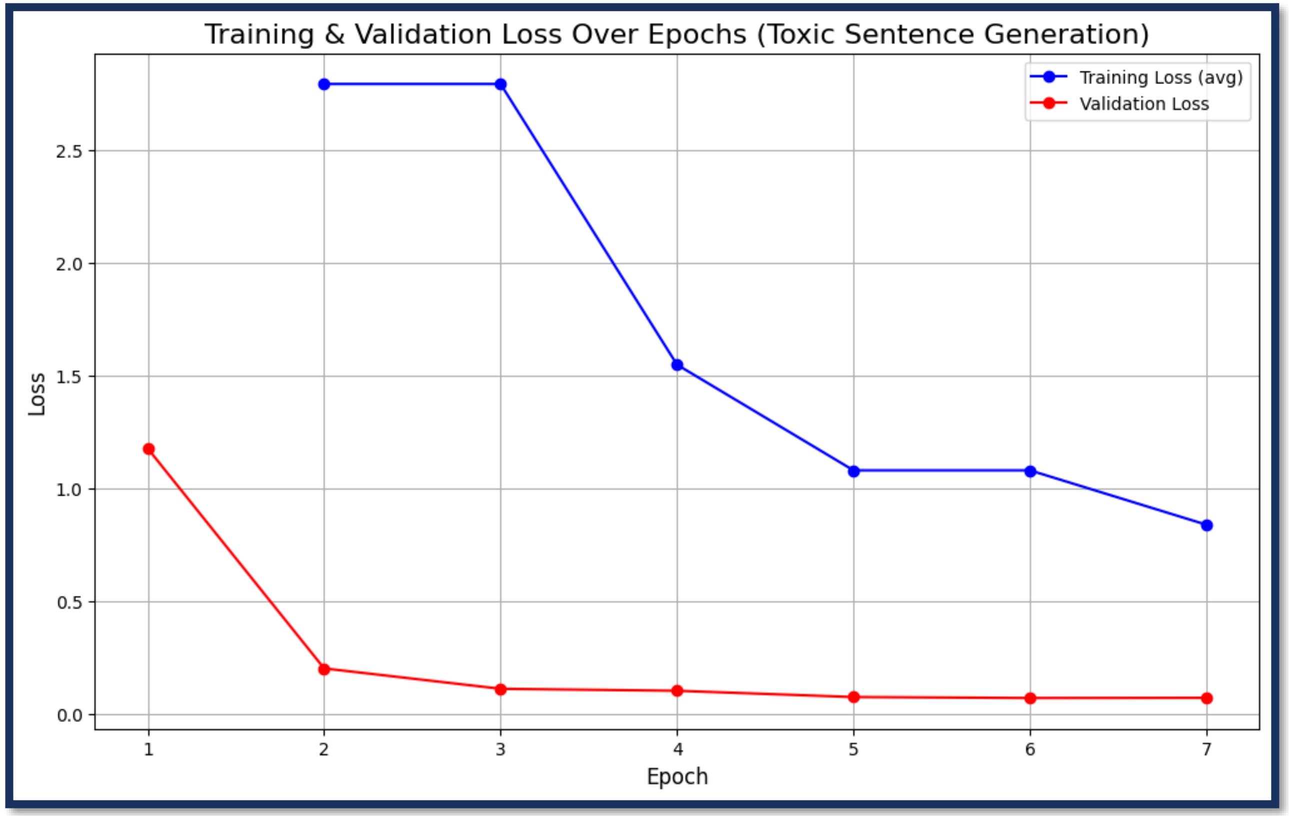
**Input Conversation:** "Sei sempre in ritardo, non mi ascolti mai.", "Non voglio più vederti, sei una delusione.", "Dobbiamo lavorare sulla nostra comunicazione."

**Conversation Category:** Litigio

**Generated Most Toxic Sentence:** 'Sei sempre in ritardo, non mi ascolti mai.'



| Metric | Average Score |
|---|---|
| BLEU | 0.305912 |
| ROUGE-1 | 0.434351 |
| ROUGE-2 | 0.369472 |
| ROUGE-L | 0.424525 |



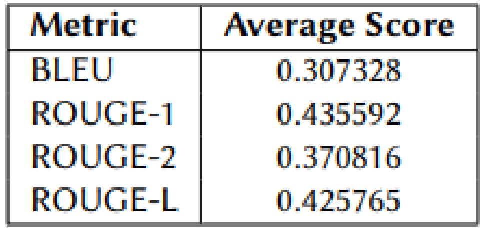
Training & Validation Loss Over Epochs (Toxic Sentence Generation)

# KEY RESULTS – T5 GENERATION

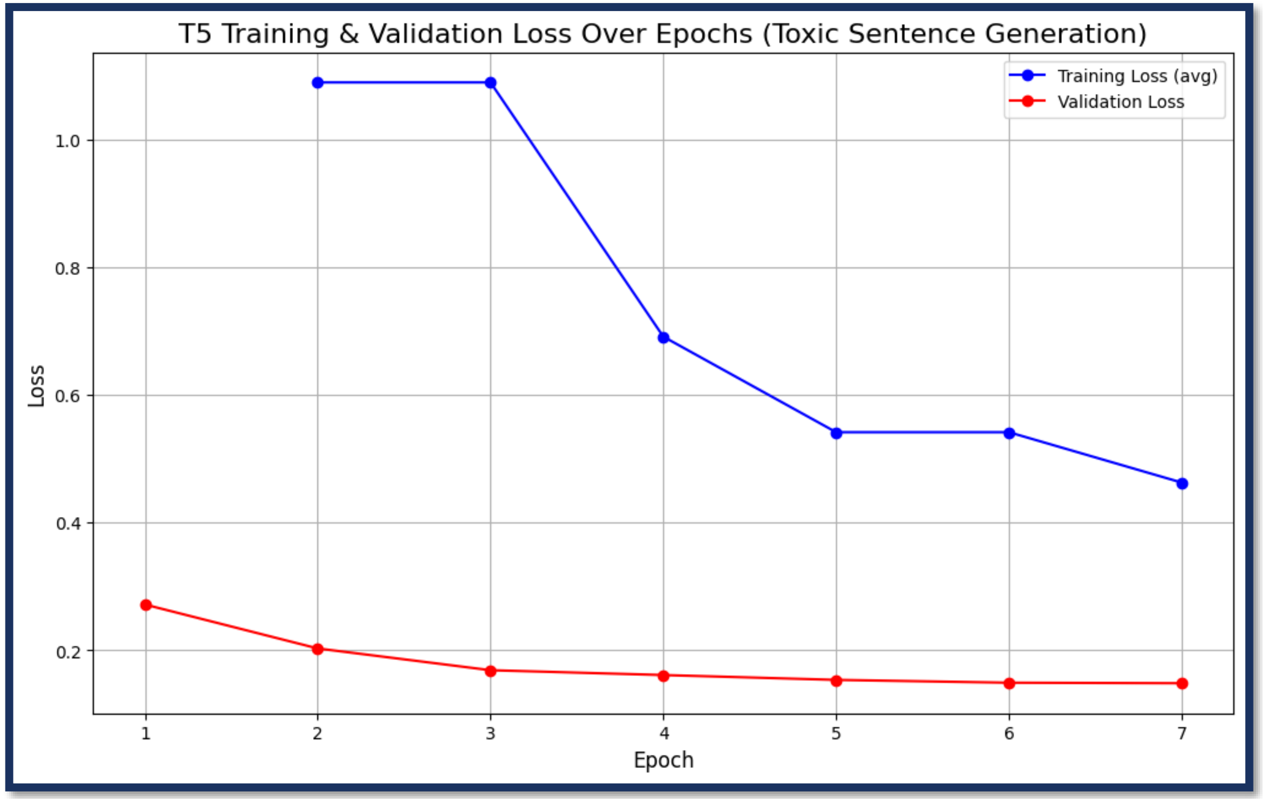
**Input Conversation:** "Sono davvero stanco di questo, non capisco perché fai così.", "Sei egoista e non pensi mai a nessuno tranne te stesso!", "Dobbiamo trovare un compromesso, questa situazione è insostenibile."

**Conversation Category:** Discussione

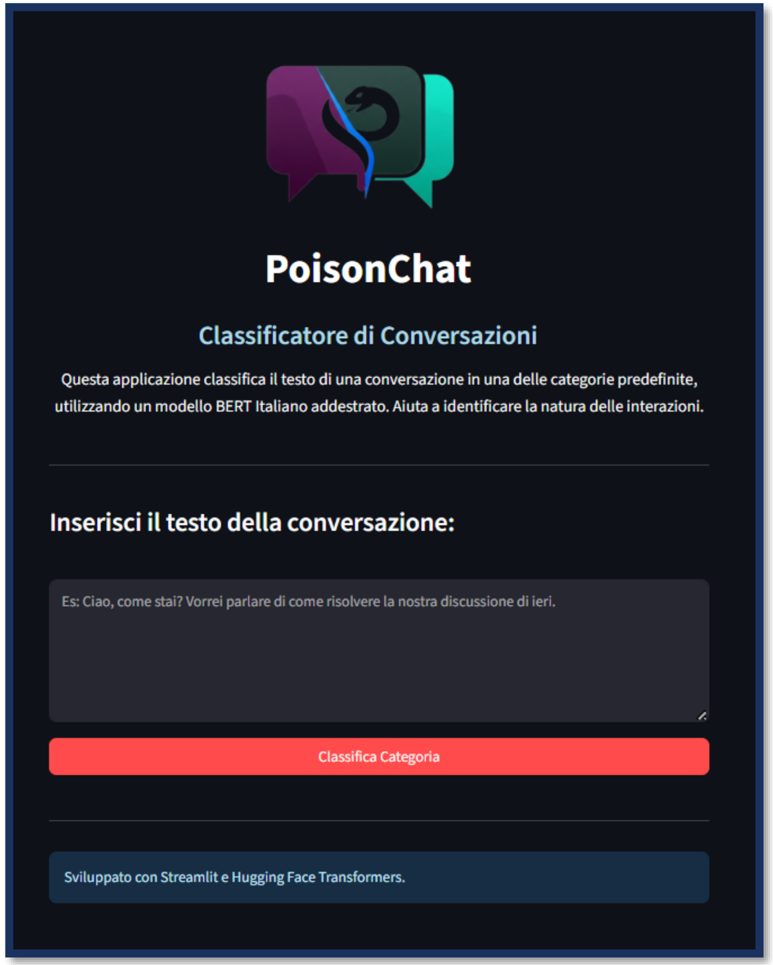
**Generated Most Toxic Sentence:** 'Sono davvero stanco di questo, non capisco perché fai così.'



| Metric | Average Score |
|---|---|
| BLEU | 0.307328 |
| ROUGE-1 | 0.435592 |
| ROUGE-2 | 0.370816 |
| ROUGE-L | 0.425765 |



# POISONCHAT - THE WEB APPLICATION



Classificatore di Conversazioni

Questa applicazione classifica il testo di una conversazione in una delle categorie predefinite, utilizzando un modello BERT Italiano addestrato. Aiuta a identificare la natura delle interazioni.

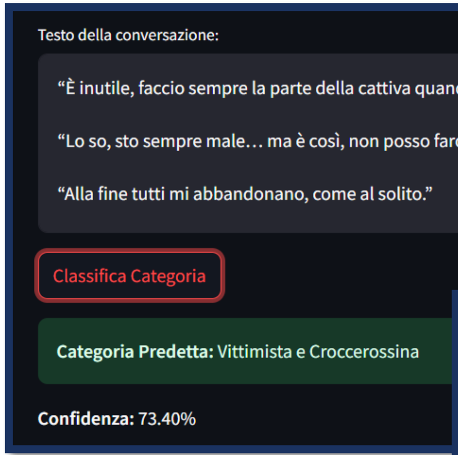
Inserisci il testo della conversazione:

Es: Ciao, come stai? Vorrei parlare di come risolvere la nostra discussione di ieri.

Sviluppato con Streamlit e Hugging Face Transformers.

- **Objective:** To provide a user-friendly and interactive real-time demonstration of the developed NLP models.
- **Technology:** Developed as a web application named 'PoisonChat' using **Streamlit**, **GitHub** and **HuggingFace**.
- **Key Features:**
  - **Text Input:** Allows users to input Italian text (e.g., chat dialogues, individual sentences).
  - **Real-time Toxicity Detection:** Provides instant feedback on the detected binary toxicity level (Toxic/Non-Toxic).
  - **Conversation Categorization:** Displays the predicted multi-class conversation category.

# POISONCHAT - EXAMPLES



Testo della conversazione:

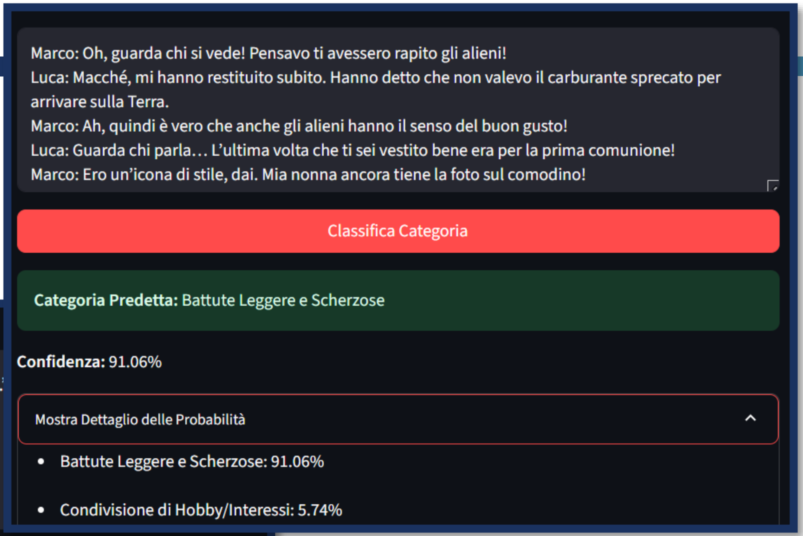
“È inutile, faccio sempre la parte della cattiva quando cerco solo di aiutare.

“Lo so, sto sempre male… ma è così, non posso farci niente.”

“Alla fine tutti mi abbandonano, come al solito.”

**Categoria Predetta:** Vittimista e Croccerossina

**Confidenza:** 73.40%

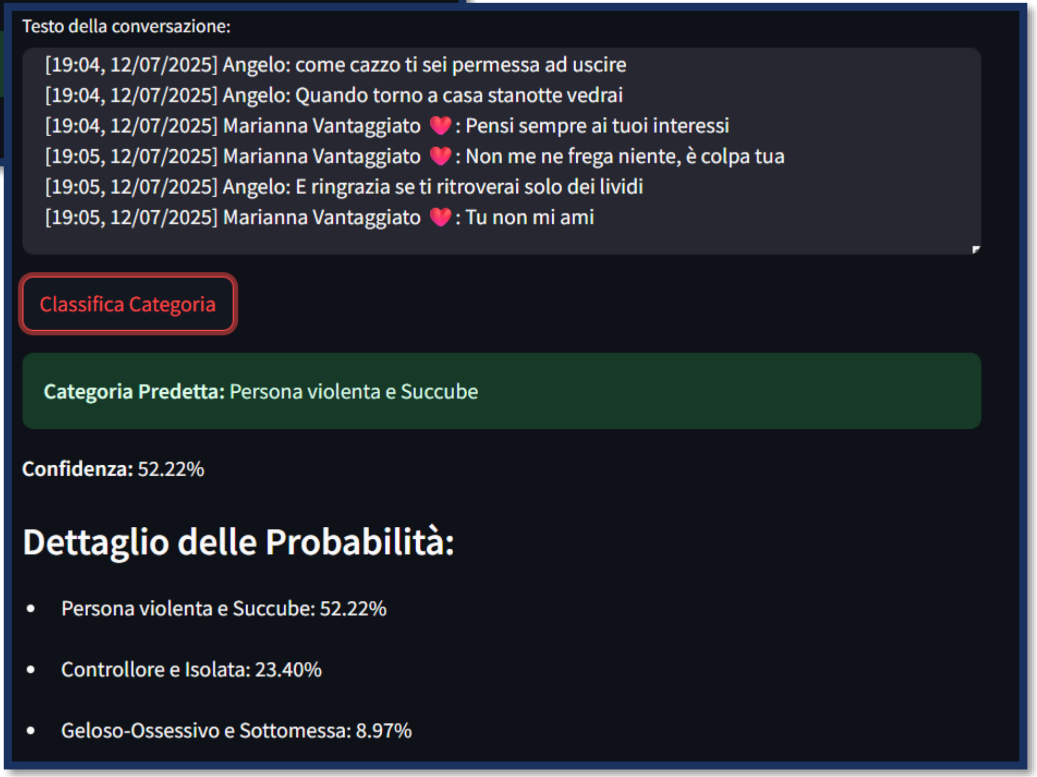


Marco: Oh, guarda chi si vede! Pensavo ti avessero rapito gli alieni!
Luca: Macché, mi hanno restituito subito. Hanno detto che non valevo il carburante sprecato per arrivare sulla Terra.
Marco: Ah, quindi è vero che anche gli alieni hanno il senso del buon gusto!
Luca: Guarda chi parla… L'ultima volta che ti sei vestito bene era per la prima comunione!
Marco: Ero un'icona di stile, dai. Mia nonna ancora tiene la foto sul comodino!

**Categoria Predetta:** Battute Leggere e Scherzose

**Confidenza:** 91.06%

- Battute Leggere e Scherzose: 91.06%
- Condivisione di Hobby/Interessi: 5.74%



Testo della conversazione:

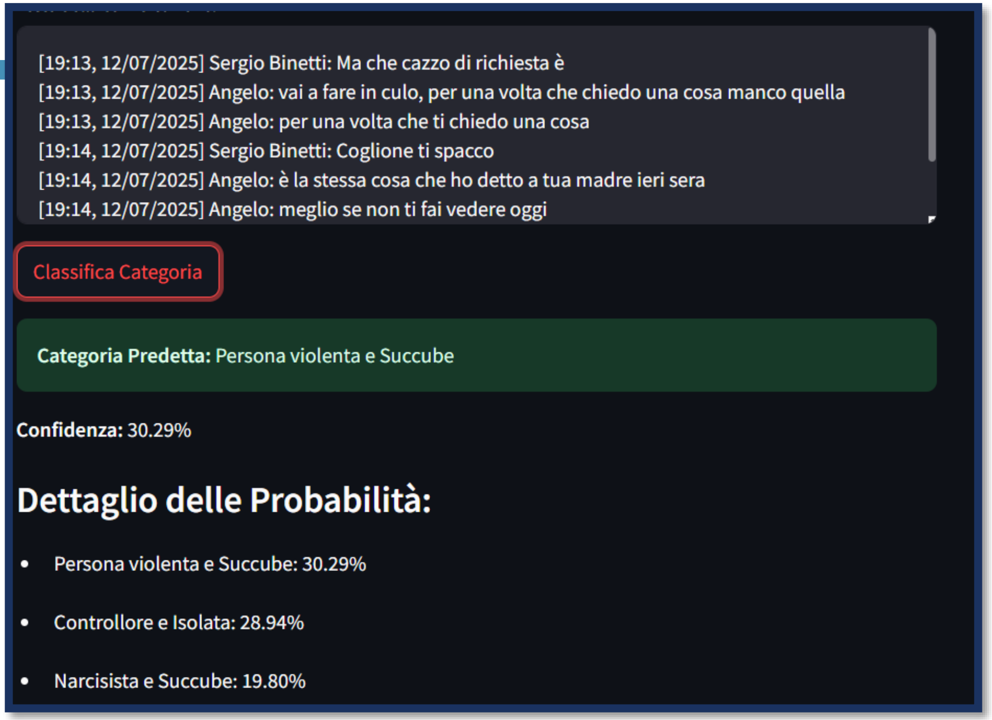
[19:04, 12/07/2025] Angelo: come cazzo ti sei permessa ad uscire
[19:04, 12/07/2025] Angelo: Quando torno a casa stanotte vedrai
[19:04, 12/07/2025] Marianna Vantaggiato ❤️: Pensi sempre ai tuoi interessi
[19:05, 12/07/2025] Marianna Vantaggiato ❤️: Non me ne frega niente, è colpa tua
[19:05, 12/07/2025] Angelo: E ringrazia se ti ritroverai solo dei lividi
[19:05, 12/07/2025] Marianna Vantaggiato ❤️: Tu non mi ami

**Categoria Predetta:** Persona violenta e Succube

**Confidenza:** 52.22%

## Dettaglio delle Probabilità:

- Persona violenta e Succube: 52.22%
- Controllore e Isolata: 23.40%
- Geloso-Ossessivo e Sottomessa: 8.97%



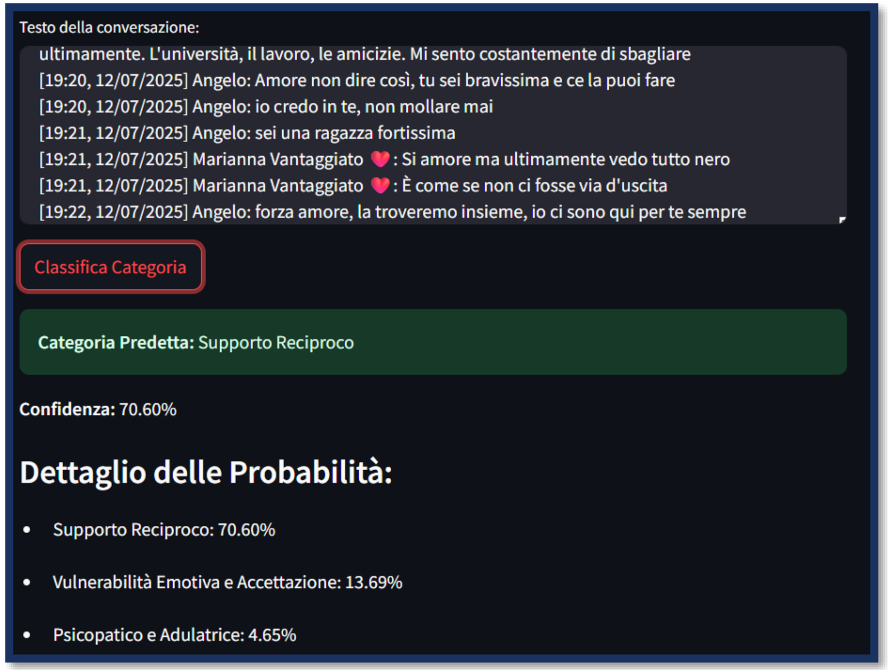
[19:13, 12/07/2025] Sergio Binetti: Ma che cazzo di richiesta è
[19:13, 12/07/2025] Angelo: vai a fare in culo, per una volta che chiedo una cosa manco quella
[19:13, 12/07/2025] Angelo: per una volta che ti chiedo una cosa
[19:14, 12/07/2025] Sergio Binetti: Coglione ti spacco
[19:14, 12/07/2025] Angelo: è la stessa cosa che ho detto a tua madre ieri sera
[19:14, 12/07/2025] Angelo: meglio se non ti fai vedere oggi

**Categoria Predetta:** Persona violenta e Succube

**Confidenza:** 30.29%

## Dettaglio delle Probabilità:

- Persona violenta e Succube: 30.29%
- Controllore e Isolata: 28.94%
- Narcisista e Succube: 19.80%



Testo della conversazione:

ultimamente. L'università, il lavoro, le amicizie. Mi sento costantemente di sbagliare
[19:20, 12/07/2025] Angelo: Amore non dire così, tu sei bravissima e ce la puoi fare
[19:20, 12/07/2025] Angelo: io credo in te, non mollare mai
[19:21, 12/07/2025] Angelo: sei una ragazza fortissima
[19:21, 12/07/2025] Marianna Vantaggiato ❤️: Si amore ma ultimamente vedo tutto nero
[19:21, 12/07/2025] Marianna Vantaggiato ❤️: È come se non ci fosse via d'uscita
[19:22, 12/07/2025] Angelo: forza amore, la troveremo insieme, io ci sono qui per te sempre

**Categoria Predetta:** Supporto Reciproco

**Confidenza:** 70.60%

## Dettaglio delle Probabilità:

- Supporto Reciproco: 70.60%
- Vulnerabilità Emotiva e Accettazione: 13.69%
- Psicopatico e Adulatrice: 4.65%

# CONCLUSIONS AND LIMITATIONS

- **Conclusions:**
  - **Superiority of Transformers:** Fine-tuned Transformer models (especially BERT) consistently outperformed traditional Machine Learning models in both discriminative (classification) and generative (toxic sentence generation) NLP tasks for Italian.
  - **Data Generation:** The ability to generate a custom corpus of non-toxic Italian dialogues using the Gemini API was crucial in addressing data scarcity and enhancing dataset quality and diversity.
  - **Practical Application:** The PoisonChat web application successfully demonstrates the practical applicability and real-time utility of the developed models.

- **Limitations:**
  - **Generative Metrics:** Reliance on BLEU and ROUGE for generative models does not fully capture semantic nuance or contextual appropriateness; human evaluation is essential for future work.
  - **Dataset Scale:** While custom and diverse, the dataset may still be limited in scale and thematic breadth compared to large-scale English datasets, potentially affecting model generalizability.
  - **Subjectivity in Annotation:** The inherent subjectivity in toxicity and conversation category annotation remains a challenge, potentially introducing noise or bias into the dataset.

# FUTURE IMPROVEMENTS

- **Enhanced Generative Model Evaluation:**
  - Integrate comprehensive human evaluation to assess semantic nuance, fluency, and contextual appropriateness of generated toxic sentences, moving beyond automated metrics like BLEU and ROUGE.
- **Dataset Expansion and Diversity:**
  - Expand the scale and thematic breadth of the dataset to enhance the generalizability and robustness of the models.
  - Explore additional sources of Italian conversational data.
- **Explore Advanced Architectures:**
  - Investigate and fine-tune newer or larger Transformer models, potentially including those specifically optimized for the Italian language, to push performance boundaries.
- **Improved Annotation Robustness:**
  - Refine annotation guidelines and conduct inter-annotator agreement studies to reduce subjectivity and improve dataset quality.
- **Addressing Class Imbalance:**
  - Implement advanced techniques (e.g., data augmentation, over/under-sampling, custom loss functions) to mitigate the impact of class imbalance, especially in tasks like "Most Toxic Sentence Classification," to improve recall for minority classes.

# THANKS

ANGELANTONIO FEDELE MUROLO
a.murolo7@studenti.uniba.it