Toxicity and Conversation Category Classification in Italian Dialogues: An Integrated NLP Approach with Fine-tuned Transformers and Web Application

Angelantonio Fedele Murolo[1,2,3]

[1]University of Bari Aldo Moro
[2]Master Degree in Computer Science - Artificial Intelligence
[3]Natural Processing Language Exam

Abstract

This paper details an integrated Natural Language Processing (NLP) system for detecting toxicity and classifying conversation categories in Italian dialogues. Addressing data scarcity, we generated a custom non-toxic dataset using the Gemini API, which was then merged with existing toxic data. Our approach evaluates both traditional Machine Learning models, including Support Vector Machines (SVM) and Logistic Regression with various feature representations, and, predominantly, fine-tuned Transformer models. Specifically, an Italian BERT model was fine-tuned for binary toxicity detection and multi-class conversation categorization. Additionally, a T5 Transformer was fine-tuned for generating or extracting the most toxic sentences. The system culminates in a Streamlit web application for real-time demonstration. Our results underscore the superior performance of fine-tuned Transformer models in handling the linguistic nuances of Italian for both discriminative and generative NLP tasks. All code, data, and models are publicly available on GitHub at https://github.com/AngeloTetro/NLP-CIPV.

Keywords

NLP, Conversational Toxicity, Text Classification, BERT, Machine Learning, Italian Language

1. Introduction and Motivation

Online communication platforms face increasing challenges from harmful content, including hate speech and general toxicity. This issue is particularly complex for the Italian language due to its linguistic nuances and limited availability of annotated datasets. Our project addresses this gap by developing a robust Natural Language Processing (NLP) system specifically for Italian conversational data.

The primary objectives of this work are to:

1. *Detect Toxicity*: Classify Italian conversations as "toxic" or "non-toxic".
2. *Classify Conversation Categories*: Categorize conversations into granular types, offering deeper insights into dialogue nature.
3. *Explore Toxic Sentence Generation/Extraction*: Investigate generative Transformer models to identify or produce the most toxic sentence within a conversation.

Our approach involves creating a balanced Italian dataset by generating non-toxic conversations using the Gemini API and combining them with existing toxic data. We employ both traditional machine learning models (e.g., SVM, Logistic Regression with TF-IDF and FastText) as baselines, and fine-tune state-of-the-art Transformer models like BERT for classification and T5 for sentence generation/extraction. The entire system is then integrated into an interactive web application built with Streamlit, providing real-time analysis capabilities. This project aims to contribute to more effective content moderation tools tailored for the Italian context.

NLP Course 2024-2025
a.murolo7@studenti.uniba.it (A. F. Murolo)
https://github.com/AngeloTetro (A. F. Murolo)

2. Related Work

The increasing prevalence of online platforms highlights a growing need for effective automated content moderation. In NLP, toxicity and hate speech identification has evolved significantly. Early approaches relied on traditional machine learning with hand-engineered features [1], but struggled with contextual nuances.

A major shift came with deep learning, especially large pre-trained Transformer models like BERT [2] and RoBERTa [3]. These models improved text understanding through contextual embeddings, boosting performance across NLP tasks, including problematic discourse detection. Recent studies emphasize the context-dependent nature of toxicity [4].

For Italian, robust content moderation tools face challenges due to limited annotated datasets. However, initiatives like EVALITA shared tasks have driven research in Italian hate speech detection, fostering resource creation and diverse methodologies, from statistical classifiers to fine-tuned multilingual Transformers [5, 6].

While existing literature often focuses on generic toxicity or specific user characteristics [7], there's a gap for comprehensive systems providing multi-faceted understanding of Italian dialogues. Our work fills this gap with an integrated NLP framework. Distinct from personality profiling, our system not only identifies toxicity but also rigorously classifies conversations into semantic categories, offering granular, actionable interpretations via a hybrid approach combining classical machine learning and advanced Transformer architectures.

3. Proposed Approach

Our project introduces an integrated Natural Language Processing (NLP) framework designed for toxicity detection and conversation category classification in Italian dialogues, culminating in a user-friendly web application for real-time demonstration. The methodology encompasses dataset creation, modeling with various machine learning and deep learning techniques, and practical deployment.

3.1. Description of the Solution and Dataset

Addressing the scarcity of Italian conversational data, we created a hybrid dataset.

- *Non-Toxic Dataset Generation*: Leveraging the Gemini API, we programmatically generated a custom corpus of diverse, healthy Italian dialogues, ensuring content purity through careful prompt engineering and safety configurations.
- *Toxic Dataset Integration*: This synthetic data was merged with an existing toxic Italian dialogue dataset (‘classification_and_explaination_toxic_conversation.csv‘), which underwent preprocessing including dialogue reformatting and extraction of the "most toxic sentence".
- *Dataset Unification and Balancing*: The two datasets were unified into ‘unified_conversation_dataset.csv‘, ensuring a balanced representation for robust model training. This dataset forms the foundation for both binary toxicity and multi-class conversation categorization.

3.2. Main Technical Details

Our approach employs a multi-model strategy, utilizing traditional machine learning baselines and advanced Transformer architectures.

3.2.1. Traditional Machine Learning Models (Baselines)

To establish performance baselines for binary toxicity classification, we trained and evaluated several classical machine learning models. These models were assessed using two feature representation techniques:

- *TF-IDF (Term Frequency-Inverse Document Frequency)*: Applied for feature extraction with Logistic Regression, Naive Bayes Classifier, LightGBM (Light Gradient Boosting Machine), and Support Vector Machine (SVM).
- *FastText Embeddings*: Pre-trained word embeddings for Italian were used to represent sentences for Logistic Regression, Naive Bayes Classifier, LightGBM, and SVM.

GridSearchCV was systematically employed across all baseline models for hyperparameter tuning.

3.2.2. Transformer-based Models (Direct Fine-tuning)

The core of our advanced capabilities relies on fine-tuned Transformer models.

- *BERT Fine-tuning for Classification*: The multi-class task involved classifying dialogues into 18 predefined categories: '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', and 'Vulnerabilità Emotiva e Accettazione'. Fine-tuning was performed using the Hugging Face 'Trainer' API, utilizing AdamW optimizer, appropriate loss functions, and early stopping.

3.2.3. Embeddings-based Classification with Transformers (BERT and RoBERTa)

- *Approach*: Contextual embeddings were extracted from pre-trained BERT ('dbmdz/bert-base-italian-uncased') and RoBERTa ('bert-base-italian-xxl-cased') models. These embeddings served as features for a Logistic Regression classifier, providing an alternative method to leverage Transformer representations for both binary and multi-class classification tasks.

3.2.4. T5 for Most Toxic Sentence Generation/Extraction

- *Model*: A T5 (Text-to-Text Transfer Transformer) model, 't5-base', was fine-tuned for the generative task of identifying or producing the most toxic sentence within a dialogue.
- *Task*: The model was trained to take a full conversation as input and generate/extract the most salient toxic sentence as per our curated dataset, enabling fine-grained toxicity analysis.

3.3. Web Application Deployment

To provide a practical demonstration, we developed a real-time web application using Streamlit.

- *Functionality*: The application allows users to input Italian conversational text and receive instant predictions regarding its conversation category. It integrates the fine-tuned BERT model, loading the tokenizer and model weights from Hugging Face Hub ('AngeloTetro/PoisonChat').
- *Accessibility*: This setup offers a user-friendly interface, making the NLP system accessible for practical content analysis and moderation demonstrations.

4. Evaluation

4.1. Evaluation Metrics

For classification tasks, we primarily used standard metrics: Accuracy, Precision, Recall, and F1-score. For multi-class classification, these metrics are typically presented with macro or weighted averages, along with per-class scores. Confusion matrices are used to visualize classifier performance by showing true positive, true negative, false positive, and false negative rates. For text generation tasks, we employed BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, which measure the similarity between generated and reference texts.

4.2. Traditional Machine Learning Model Performance

The baseline models were evaluated for binary toxicity classification using TF-IDF and FastText embeddings.

Table 1
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

4.3. Transformer-based Model Performance

4.3.1. BERT Fine-tuning for Multi-class Conversation Classification

The fine-tuned BERT model achieved an overall accuracy of **0.8232** on the test set for conversation category classification. Detailed per-class performance is illustrated by the confusion matrix as it follows:.

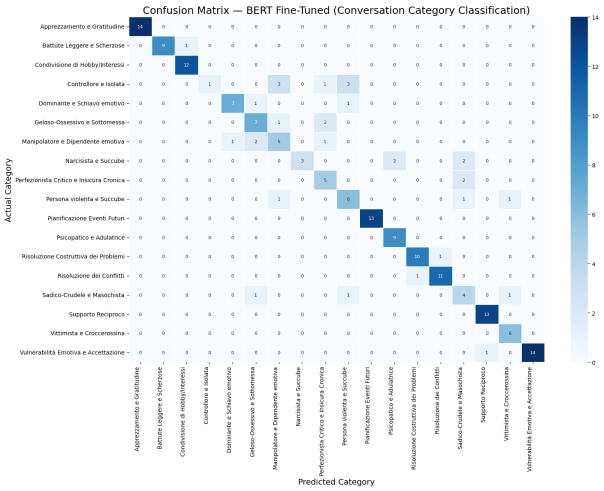

Figure 1: Training and Validation Loss for BERT Fine-tuned Multi-class Conversation Classification.generalization.

4.3.2. Embeddings-based Classification (BERT and RoBERTa)

The performance of BERT and RoBERTa embeddings combined with Logistic Regression for conversation category classification is summarized below.

Table 2
Overall Accuracy for Embeddings-based Classification Models

Model	Test Set Accuracy
BERT Embeddings + Logistic Regression	0.7735
RoBERTa Embeddings	0.76

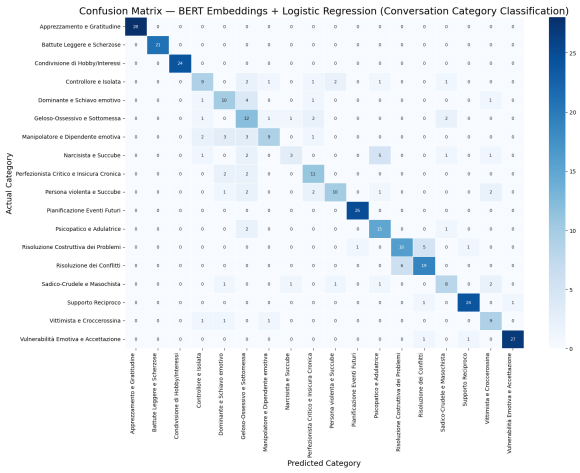

Figure 2: Confusion Matrix for BERT (Categories)

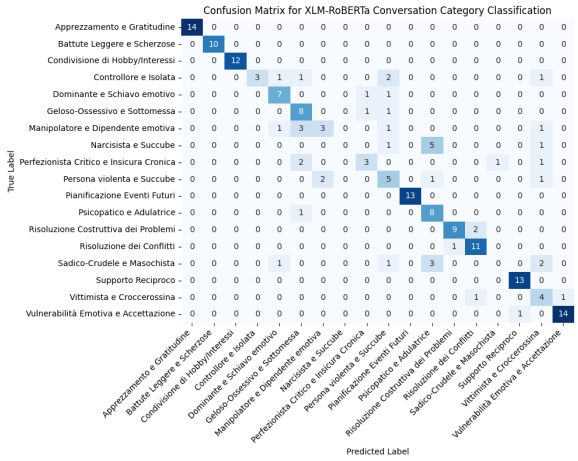

Figure 3: Confusion Matrix for RoBERTa (Categories)

4.3.3. BERT for Most Toxic Sentence Classification

A BERT fine-tuned model was also used to classify individual sentences as "Toxic Sentence" or "Non-Toxic Sentence", serving to identify the most toxic sentence within a conversation.

Table 3
Performance Metrics for BERT Most Toxic Sentence Classifier

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
Macro Avg	0.46	0.50	0.48	597
Weighted Avg	0.84	0.92	0.88	597

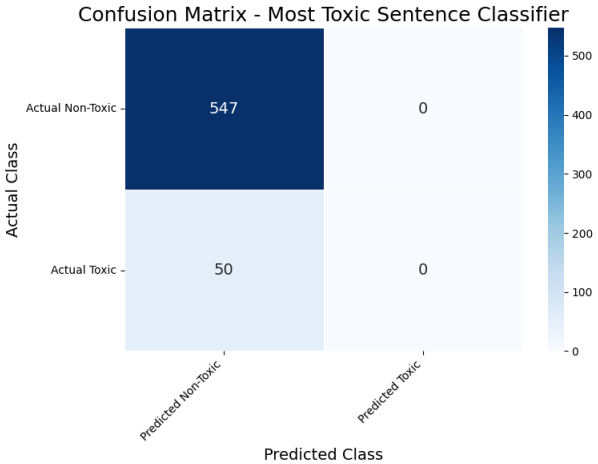

Figure 4: Confusion Matrix for BERT Most Toxic Sentence Classifier

Example prediction on new data:

- 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)

4.3.4. Transformer-based Models for Most Toxic Sentence Generation (BART and T5)

The BART model was fine-tuned for generating or extracting the most toxic sentence from a given conversation. Similarly, the T5 model was also fine-tuned for this task. The average scores on the test set for both models are presented below:

Table 4
Comparative BLEU and ROUGE Scores for BART and T5 Most Toxic Sentence Generation

Table 5
BART Scores

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

Table 6
T5 Scores

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

Example generation with T5 on new data:

- 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."]
- Generated Most Toxic Sentence (T5): 'Sono davvero stanco di questo, non capisco perché fai così.'

4.4. Discussion and Results

This section analyzes the performance of traditional machine learning and Transformer models for Italian toxicity detection and conversation categorization, highlighting key strengths, limitations, and practical implications.

For binary toxicity classification, traditional models, particularly SVM with TF-IDF features, showed exceptional performance (1.0 accuracy, precision, recall, F1-score for "Toxic" class, Table 1). This near-perfect result, while indicative of strong data separability, warrants caution as it could suggest overfitting or an overly simplistic dataset for this task. Logistic Regression with TF-IDF also performed well, outperforming FastText embeddings, implying TF-IDF was more effective for this specific problem.

In multi-class conversation classification, the fine-tuned BERT model achieved the highest accuracy (0.8232). It excelled in clear categories (e.g., "Appreciation and Gratitude," F1-score 1.00) but struggled with ambiguous or under-represented ones (e.g., "Controller and Isolated," F1-score 0.22). Embedding-only approaches (BERT Embeddings + Logistic Regression at 0.7735 accuracy, RoBERTa Embeddings at 0.76 accuracy, Table 2) performed lower, emphasizing the benefit of fine-tuning for complex tasks.

Regarding most toxic sentence generation, BART and T5 demonstrated decent performance, with T5 slightly better. While metrics indicate lexical overlap, they don't fully capture semantic accuracy or fluency. Generating precise toxic sentences remains challenging, suggesting a need for more advanced techniques or evaluation.

In conclusion, Transformer models are promising for Italian NLP, especially with fine-tuning for complex classification. However, managing class imbalance is crucial. For generation, BART and T5 provide a solid foundation. These models offer powerful tools for analyzing Italian language nuances, as practically demonstrated by the web application.

4.5. Web Application

To provide a practical demonstration and a user-friendly interface for real-time analysis of Italian conversations, the entire system has been encapsulated within an interactive web application called **PoisonChat**, built with Streamlit. This application, accessible at https://poisonchat.streamlit.app/, allows users to input conversational text and receive classifications for toxicity and conversation categories.

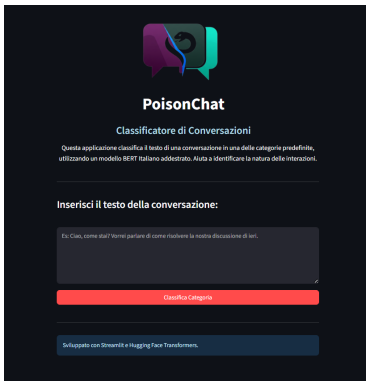

Figure 5: WebApp PoisonChat, conversation classifier

5. Conclusions and Limitations

This work presented an integrated NLP framework for Italian conversation analysis, covering toxicity detection, multi-class categorization, and most toxic sentence generation, demonstrated via a web application.

Our findings showed traditional models excelled at binary toxicity, though perfect scores warrant cautious interpretation regarding potential data simplicity or overfitting. Fine-tuned Transformers (BERT) proved superior for multi-class classification, despite varying performance across categories, highlighting challenges with ambiguous classes.

A critical limitation was the classification of the most toxic sentence, where severe class imbalance led to complete failure in identifying the minority toxic class. This underscores the vital need for robust imbalance handling. For generation, BART and T5 showed decent performance, but current metrics don't fully capture semantic quality, indicating areas for future work in evaluation and model enhancement.

References

[1] A. Schmidt, M. Wiegand, A survey on hate speech detection, Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2017) (2017) 1–10.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (2019) 4171–4186.

[3] Y. Liu, X. Gao, J. Han, M. Sun, Y. Wu, Z. Liu, L. Cui, N. Xu, Y. Zhang, Y. Ding, et al., Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[4] D. Xenos, I. Nikitina, R. Kumar, H. Ma, Toxicity in dialogue: Context matters, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), 2021, pp. 5777–5787.

[5] C. Bosco, F. Tamburini, T. Caselli, F. Dell'Orletta, A. Lenci, Evalita 2018: Overview of the 6th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018), 2018, pp. 1–14.

[6] F. Poletto, M. Sanguinetti, C. Bosco, Resources and tools for hate speech detection in italian: A survey, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2021), 2021, pp. 6706–6715.

[7] F. Mairesse, M. A. Walker, R. Mihalcea, J. Wiebe, Linguistic correlates of personality in conversation, in: Proceedings of the HLT-NAACL 2007 Workshop on Bridging Disciplines: Computational Linguistics and Biomedical Informatics, 2007, pp. 25–32.