CIPV - Cyber Intimate Partner Violence

Filippo di Gravina[1]

Abstract
Intimate partner violence refers to behaviour within an intimate relationship that causes physical, sexual or psychological harm, including acts of physical aggression, sexual coercion, psychological abuse and controlling tendencies. This definition covers violence by both current and former spouses and partners. Cyber IPV is a technology-mediated form of violence. To address this problem, and to orient prevention effort, the proposed work analyzes the various forms of violence, classifying the various personality types involved in the conversations.

Keywords
C-IPV Classification, Sentence Transformer, Natural Language Processing, Multi-class classification

1. Introduction and Motivations

Cyber Intimate Partner Violence (C-IPV) is a form of abuse directed towards one's partner via social media and digital platforms. C-IPV can come in various forms, and one of the major aspects in the prevention phase of this kind of violence is the detection of the different types of abuse. Doing such in a relationship helps to find, monitor and counter the emotional, verbal and physical violence that the perpetrator inflicts upon the victim. In this study, the main task is to identify the type of toxicity in a toxic conversation given a dataset comprised of conversations between a perpetrator and a victim.

2. Related Work

Classification of Cyber Intimate Partner Violence has recently been studied by the paper LLaMAntino against Cyber Intimate Partner Violence [1]. The authors used LLMs with different prompt techniques to explain toxicity in conversations.

Specifically, they answered the two research questions:

- is the model able to recognize toxic sentences?
- are the explanations provided with 2-shot prompting similar to the "gold standard" provided by experts?

The experiments lead to the following results:

- few-shot prompting always outperforms zero-shot prompting
- even with few-shot prompting, the LLM learns to provide good explanations and learns to detect the presence or the absence of the toxicity

NLP Course 2024-2025
f.digravina2@studenti.uniba.it (F. d. Gravina)
https://github.com/fdigravina/ (F. d. Gravina)

3. Proposed Approach

3.1. Description of the solution and dataset

The dataset used in the task is provided by the study by Vito Nicola Losavio called IDaToC: An Italian Dataset of Toxic Conversation. [2]. This dataset provides a thousand of examples of synthetic conversations generated by large language models, along with the explanation of the toxicity, a binary feature representing if the conversation is toxic or not, and a column representing the roles of the two people involved in the conversation. This column contains 10 different values:

- 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
- Psicopatico e Adulatrice
- Sadico-Crudele e Masochista
- Vittimista e Crocerossina

Various methods have been tested to find the best solution for this multi-class classification. Based on the related works, the chosen baseline is to evaluate the accuracy of a LLM through zero-shot prompting classification and few-shot prompting classification. All the tests are made using prompt in Italian and in English.

The zero-shot prompting classification led to very poor results, giving an accuracy score slightly higher than a random guess. Various LLMs have been tested, with the best result in this phase equal to a poor 14% of accuracy. The LLM that gave the best result was meta-llama/llama-4-scout-17b-16e-instruct.

Few-shot prompting classification gave slightly better results, leading to a 30% accuracy in the best case. These results have been improved by adding some additional information, such as the result of the keyword extraction algorithm, the topic labeling output and giving in input the explanation of the violence available in the dataset.

The keyword extraction algorithm is used to find the most relevant words in the conversation, in order to focus more attention on the relevant parts of the sentences.

For the topic labeling task, the tested approaches are the LSA method (Latent Semantic Analysis) and the LDA method (Latent Dirichlet Allocation). The LDA method gave the best results.

The results of the LDA are printed on an interactive web page for the entire training set and for each single test class (available on the image folder of the case study), giving the top 30 most salient terms and printing a graph of the found clusters. Hovering with the mouse on the clusters highlights the relevance of the words for that specific cluster.

Even with few-shot prompting and additional information, such as keyword extraction, topic labeling and the explanation, the best result only achieved 45% of accuracy. Even in this case, the LLM that gave the best result was meta-llama/llama-4-scout-17b-16e-instruct. Another limitation of this approach is the finite daily number of requests available on the APIs of public LLMs.

The final approach, leading to the best results, uses sequence transformers.

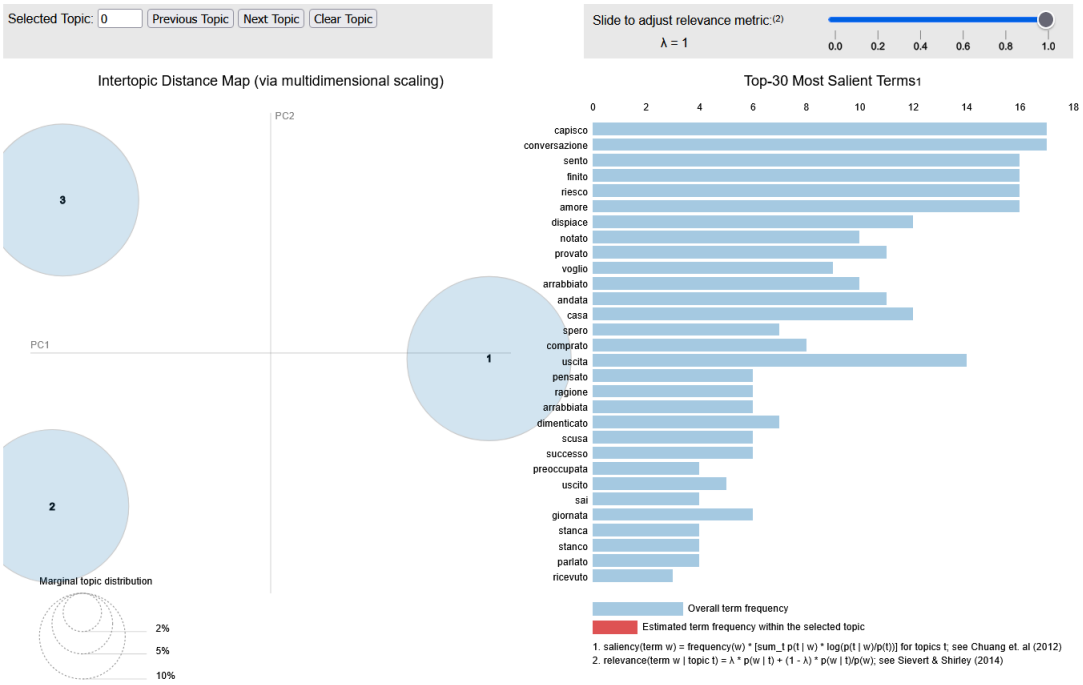

Figure 1: Results of the Latent Dirichlet Allocation

3.2. Main technical details

Sequence transformers are used to encode the input sentences and give these encodings in input to machine learning models that classify these new instances. About ten sequence transformers were tested, and the best sequence transformer turned out to be the model intfloat/multilingual-e5-large [3]. This model is available on HuggingFace.

Five machine learning models were tested in this phase:

- K-NN
- Multi Layer Perceptron
- Random Forest
- Support Vector Machine
- Logistic regressor

Every model has been heavily tested on hundreds of parameter combinations using a grid search and a 10-fold cross-validation.

The best tested model for this multi-class classification turned out to be the Support Vector Machine, leading to nearly 65% of accuracy.

3.3. Other information useful to replicate the approach

The project is made in Python. A requirements.txt is provided in the project folder. The only other requirement is an internet connection.

4. Evaluation

The grid search and cross validation helped to find the best hyperparameters. The best found hyperparameters for the SVM are:

- 'C': 20
- 'decision_function_shape': 'ovo'
- 'gamma': 'auto'
- 'kernel': 'sigmoid'
- 'max_iter': -1
- 'tol': 0.0001
- 'verbose': False

Corresponding to the following results:

- SVM accuracy: 0.62
- Micro precision average: 0.62
- Macro precision average: 0.63
- Micro recall average: 0.62
- Macro recall average: 0.62
- Micro f1 average: 0.62
- Macro f1 average: 0.62

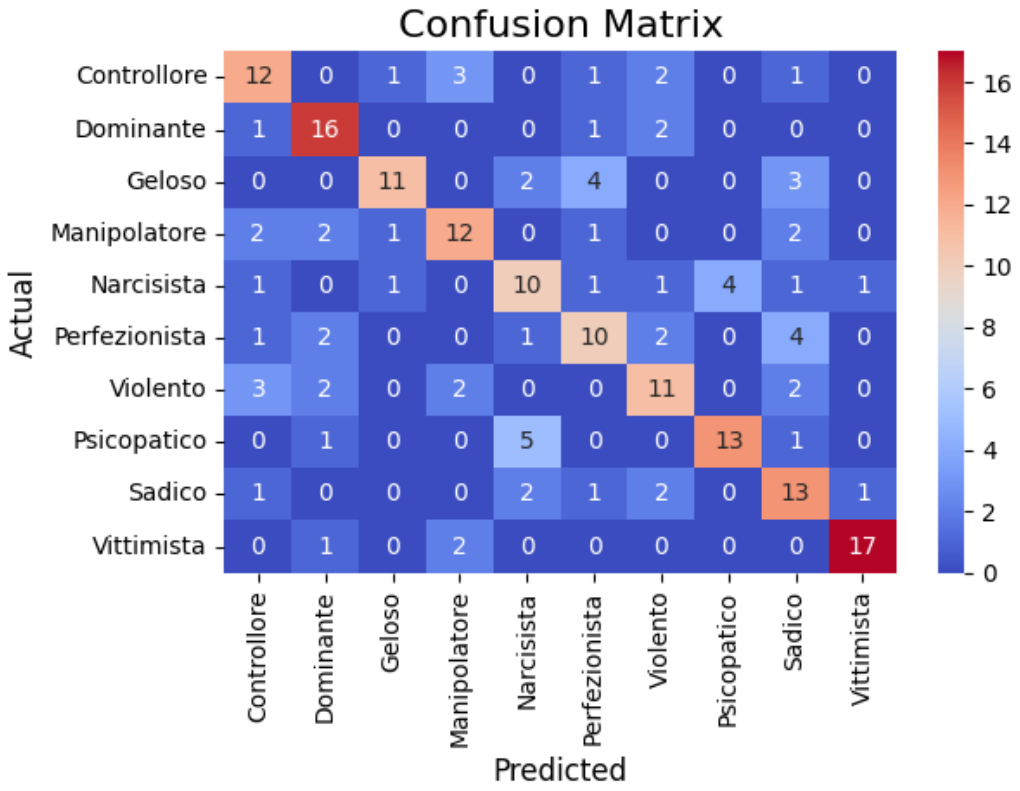

Figure 2: Confusion matrix

Micro-averaging and Macro-averaging of the metrics gave very similar results, pointing out the stability of the model. The confusion matrix highlights the fact that only a few classes are sometimes confused, for example the classes "Psicopatico e Adulatrice" and the class "Narcisista e Succube".

4.1. Robustness of the results

Other than the previously mentioned metrics, two other metrics are computed in the training phase. These two scores are computed using the grid search, and specifically they are the mean test accuracy and the standard deviation of the test accuracy in the cross validation step for the best model. The very low standard deviation value (0.03) indicates that the model is consistent and reliable in its predictions.

5. Conclusions and Limitations

The limited size of the dataset and the little availability of fine-tuned model on Italian languages, and specifically to Italian hate-speech detection models, lead to initial poor results. The proposed method gives a significantly higher accuracy, but this result is still highly improvable.

With a larger corpus of data, fine-tuning becomes a powerful option and can improve the results obtained by the Support Vector Machine, that is computationally expensive to run on datasets that contain more than tens of thousands of instances.

References

[1] P. Basile, M. Degemmis, M. Polignano, G. Semeraro, L. Siciliani, V. Tamburrano, F. Battista, R. Scardigno, LLaMAntino against cyber intimate partner violence, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 52–58. URL: https://aclanthology.org/2024.clicit-1.7/.

[2] V. N. Losavio, Idatoc: An italian dataset of toxic conversation, 2024.

[3] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, 2024. URL: https://arxiv.org/abs/2402.05672. arXiv:2402.05672.