Against Toxicity LLaMA 3

Vito Nicola Losavio[1,*,†]

[1]MSc student of University of Bari Aldo Moro

Abstract

ATTENTION: **In this work you will find toxic language.**
This project addresses the problem of toxic online language in the Italian language, developing deep learning and LLM models to classify toxic phrases and generate explanations about their toxicity.
The proposed methodology involves fine-tuning a pre-trained LLM on a curated dataset of Italian toxic phrases, including artificially generated examples and cases manually labeled by psychologists.
The preliminary results show that the developed models achieve high levels of accuracy in classifying the type of violence and aggressive communication, with average values exceeding 98%. Furthermore, the models are able to generate clear and contextual explanations about the toxicity of the phrases.
This work represents an important step forward in the fight against toxic online language, providing effective tools to identify and understand harmful behaviors in Italian communication.

Keywords

Toxic Language, Multiclass classification, Deep learning, Quntizzated LLaMa 3, LoRa, Fine tuning

1. Introduction and Motivations

The recognition of toxic behavior in relationships, or more generally the identification of manipulative and toxic phrases in the Italian language through the use of artificial intelligence, is becoming a significant challenge that researchers are striving to address. Toxic language, often characterized by subtle manipulation, verbal abuse, and emotionally harmful content, can severely impact mental health and interpersonal dynamics. Therefore, developing effective tools to detect and explain toxic language is critical for promoting healthier communication and preventing toxic relationships.

In this work, we will employ deep learning techniques in conjunction with Large Language Models (LLMs) to classify toxic phrases and generate explanations for why these phrases are considered toxic. Deep learning models, particularly those based on transformer architectures, have demonstrated exceptional performance in various natural language processing tasks, including sentiment analysis and toxicity detection. By leveraging these advanced models, we aim to accurately identify toxic language in Italian.

Our methodology involves fine-tuning pre-trained LLMs, specifically a Quantized version of LLaMa 3, on a carefully curated dataset of Italian toxic phrases. This dataset comprises both artificially generated phrases and manually labeled examples. The artificial dataset was created using generative models to simulate a wide range of toxic language scenarios, ensuring diversity and comprehensiveness. In contrast, the manually labeled dataset includes real-world examples of toxic phrases, annotated by psychologists to provide high-quality, contextually relevant data.

With this work we will go **1)** Analyze the state of the art about the recognition of hate text, **2)** We will talk about the construction of the dataset and finally **3)** About what has been done to classify sentences and generate an explanation.

NLP Course 2023-2024
*Corresponding author.
†These authors contributed equally.
v.losavio5@studenti.uniba.it (V. N. Losavio)
https://github.com/VitoNicolaLosavio/NLP (V. N. Losavio)
0000-0002-0877-7063 (V. N. Losavio)

2. Related Work

The classification of toxic comments and messages has gained significant attention in the field of natural language processing (NLP) due to the growing prevalence of online abuse and its detrimental effects on individuals and communities. Advances in machine learning and deep learning have driven substantial progress in this area, with a focus on developing robust and accurate models capable of identifying and mitigating toxic content.

The proposed approaches vary from simple models to increasingly complex models that include Deep Learning models and transformers. For example, Kumar et al [1] proposed an MCGiGRU for the recognition of toxic comments in a multilabel setting or for example Garlapati et al [2] provide a novel usage of Natural Languade Processing approach to classify the type of toxicity in comments.

Some authors such as Zhang et al. [3] or previously Wang et al [4] have proposed the use of LLM for the classification of toxic contents.

And finaly Polignano et al [5] propose an italian version of BERT that significantly improves over the state of art for hate to speech detection benchmarks in the Italian language.

With this we have reached the prefixed point number **1)**.

3. Proposed Approach

3.1. Description of the Dataset

For the development of this work, we started from a dataset containing only the sentences and the names of the different features, but not the values of the above. To label the dataset, for this work he collaborated with a master's student Dalena Martina of Community psychology and promotion of change and social welfare of the University of Padua. According to the Codebook we have labeled the dataset.

The features that we went to manually label are the following:

- **Type of physical violence**[6] this label can have the following values: *Violenza fisica*, *Violenza sessuale*, *Violenza/Aggressività psicologica* o*Stalking, monitoraggio e controllo*
- **Type of Cyberviolence** [7] this label can have the following value: *Violenza sessuale cyber*, *Violenza/aggressività psicologica cyber* and *Cyber stalking, monitoring, and control*
- **Aggressive communication**[8]: this can have just the values *YES* or *NO*
- **Type of aggressive communication**[9] that can have the following values: *Maledizioni*, *Ridicolizzazione/derisione*, *Parolacce*, *Minaccia* and *Attacchi (on competence, character, background, physical appearance)*

For each sentence, we can identify even more toxic behaviors, therefore we can find more features linked with a comma.

In addition to the features provided by the psychologists who made the dataset, it was thought to add an additional feature called **Description** which contains the explanation for which a sentence is toxic.

For example, in the table we have a toxic phrase labeled with the help of the psychologist.

3.2. Pre-Processing

This work is structured in two phases, the first is to classify the sentence identifying the type of violence and the type of aggressive communication if present, therefore, considering that the sentence cannot be at the same time a toxic sentence of type: Type of physical violence and Type of Cyberviolence, the first feature target named **Type of violence** was created, which can take all the possible values of the previously named labels.

And the second feature target corresponds to **Type of Aggressive Communication**, which instead of null values, contains the value: *"Nessuna"* as there is no aggressive form of communication.

sentence	Type of physical violence	Type of Cyberviolence	Aggressive communication (YES/NO)	Type of aggressive communication	Description
Il mio ragazzo dice che le donne non servono a niente.	Violenza/ Aggressività psicologica		YES	Ridicolizzazione/ derisione	Questa frase è un esempio di violenza psicologica in quanto mira a ridicolizzare e sminuire un intero gruppo di persone

Table 1
Dataset labeled by psychologist, when labeling the dataset, as we can see, the phrase can be identified with a single type of violence, therefore the labels can be empty. Also for phrases that do not contain an aggressive form of communication, they will have the label type of aggressive communication empty.

Considering the previous steps, we find ourselves in a multiclass and multilabel task, in this work we went to divide the problem of multilabel going to perform two separate multiclass classification tasks.

That said, we proceed to analyze the two separate feature targets.

3.2.1. Type of Violence and Type of Aggressive Communication

For the feature targets, we have unbalanced datasets because during the labeling and the visualization of the distributions of the data, we noticed that in some categories there were only a few examples. To obviate this, oversampling was carried out, but in a controlled manner. That is, no use was made of classic oversampling libraries, but ChatGPT [10] was used for the generation of examples similar to those given as input, and manual labeling was conducted, the labeling wad just for the category of referement, so after that, we obtained two balanced datasets, one for the type of violence and one for the aggressive communication.

The dataset with oversampling containing the features regarding the type of violence will also be used to perform the first fine tuning, as a result, the automatic labelling of the type of aggressive communication will be carried out and also the similarity will be calculated, as we will see later.

3.2.2. Compute the sentence similarity

As already announced, we have used the dataset with oversampling of the type of violence, also for the generation of the explanation through LLM, but now we will see a crucial part of this work.

In the automatic dataset labeling, we used a very basic deep learning model, which we will talk about later, but for the explanation of why the phrase is toxic, not having the opportunity to further question psychologists, this process is automated.

To automate this process, first of all, BERT [11] was used for the generation of phrase embendding, after that we went to calculate the similarity of the cosine between the sentences belonging to the same category and we went to take the phrase on which the similarity of the cosine turned out to be the greatest, doing this but we add a little error, as it goes to take the first sentence with maximum similarity, we may miss another sentence also similar, but more precise.

3.2.3. Data Augmentation

For the explanation generation, we used both the just discussed dataset and a totally artificial dataset to generate the aforementioned dataset,a quantized version of LLaMa 3 Instruct was used, this model can be less precise for is quantizzation [12], but we choose them for the limited computational resources

Starting from the dataset discussed above, with a prompt, defined by different tests, a similar sentence was generated for each example within the dataset, for this experiment we iterated this process twice generating a total of 1000 examples.

For automation, two models were used for self-labelling and for explanation, the similarity method described above was used.

With this we have reached the prefixed point number **2)**.

3.3. Models

In this section we will talk about the models and techniques used in the two tasks

3.3.1. Classification

In the development of the classification, we tried to use simple and basic models such as SVM and MLP, with a repeated stratified k-fold cross validation, but the above did not turn out to be totally suitable for the task.

Later we used very basic models for classification, which contain 2 Dense layers with Relu activation and a classification layer containing a Dense with the number of possible classes and a Softmax for classification.

Before we go to classify, we apply the CountVectorizer on sentences and the LabelBinarizer on feature target, to move to a domain where deep learning models are applicable.

Given the limited number of examples and the distribution of data that was not totally fair between classes, a stratified k-fold cross validation with 10 folds was used, these models as we will see in the final evaluation have performed very well with a precision and recall average of 98%.

3.3.2. Generation

As mentioned above, LLaMa 3 quantized was utilized for generating explanations with a priori knowledge following instruction tuning. Subsequently, two steps of fine-tuning were performed to simulate continual learning [13].

For the fine tuning we used the first dataset already discussed labeled under human supervision, and for the second instead we went to use an artificially generated dataset.

To fine tune with a device with limited resources, PEFT[14] was used, in combination with LoRa[15], the above techniques have allowed us to fine tune the model easily generalizing the problem that we want to solve with the first fine tuning, and specializing it through the second fine tuning. The second fine tuning step was performed only to verify that the model continued to learn and generated better explanations.

To carry out these steps, we used unsloth which allowed us to train the model quickly and to use the methodologies described above.

With this we have reached the prefixed point number **3)**.

4. Evaluation

Let us now describe how the models were evaluated

4.1. Evaluation of Classification models

For the classification task, we used metrics already known as precision, recall and accuracy, going to evaluate the precision of simple models (2), we go to find that the models have a precision of 0.89 for the task of classification of the type of violence and of 0.81 for the task of classification of the aggressive communication type.

These results being obtained through an iterative process of cross validation, have been scaled by a factor of 100 since the stratified k-fold cross validation was repeated 10 times with 10 fold.

Ideally these results would be good, but for the success of the task, you went to maximize as much as possible precision through deep models, They found for the type of violence a precision equal to 0.99 and for the type of aggressive communication equal to 0.986. For these metrics the result was scaled by a factor of 10 as only the stratified k-fold with 10 fold.

Type of violence		
	SVM	Deep learning model
Accuracy	0.88	0.98
Precision	0.89	0.99
Recall	0.88	0.99
Type of aggressive comunication		
	MLPClassifier	Deep learning model
Accuracy	0.78	0.99
Precision	0.81	0.99
Recall	0.79	0.99

Table 2
Result classification

4.2. Evaluation of generation

To evaluate the generation of explanations by LLaMa with or without fine tuning, we chose to make a not very rigorous evaluation, analyzing what it generated and what is found to meet the above objectives: the explanation must be in Italian and be as general as possible. We chose this type of assessment because the original model has been optimized for English and because the model does not have to give an accurate explanation, but must generate an alarm bell to the user.

The model has been evaluated with examples included in the dataset and out of domain examples to assess how clear it is with examples never seen.

For example, if we consider the phrase: *"Il mio ragazzo mi dice che sono una troia perché pubblico foto in bikini."* , and let's ask the different models to generate an explanation as to why it is toxic, considering also the label that have been automatically calculated, the basic model llama 3 will generate the following explanation: *"This response is in Italian, in the third person, and explains why the sentence is toxic. The response also considers the type of violence and the type of aggressive communication. Questa frase è tossica perché rappresenta un esempio di violenza verbale che utilizza parolacce e ridicolizzazione/derisione per denigrare e umiliare l'altra persona, creando un ambiente di aggressività e mancanza di rispetto."*

The above explanation respects point 2 of the objectives, but does not fully respect objective 1 because the model generates a first part in English and then move to Italian. The model obtained from the fine tuning of about 500 examples and the final model, generate the following response:
"Questa frase è tossica perché rappresenta un esempio di violenza cyber-stalking, monitoraggio e controllo, che utilizza parolacce e ridicolizzazione/derisione per denigrare e umiliare l'altra persona, creando un ambiente di aggressività e mancanza di rispetto. La frase è un esempio di comunicazione aggressiva che mira a controllare e limitare la libertà dell'altra persona"

This answer fully respects the designated task, but we puntualize that this initial phrase is a phrase contained in the dataset.

Let's now analyze sentences out of domain to see if the models developed are suitable for contexts never seen.

Let us now consider the sentence: *"Mi ha preso il telefono con la forza e non me lo ha voluto restituire"*. The basic model llama generates the following answer: *"This response is in Italian, in the third person, and explains why the sentence is toxic. The response identifies the type of violence and aggressive communication used in the sentence. The sentence 'Mi ha preso il telefono con la forza e non me lo ha voluto restituire' is toxic because it represents an example of psychological violence, where the aggressor uses physical force to take away someone's property and refuses to return it, creating a sense of powerlessness and disrespect."*. In this context, the generated sentence does not comply with the first objective. On the other hand, the models developed do not generate a correct and consistent response as they produce the following explanation: *"I cannot write a response that promotes or justifies violence or aggressive communication. Is there anything else I can help you with?"*, which indicates a poor interpretation of the prompt.

Finally, let's consider the above sentence:*"Mi ha preso con forza e mi ha portata via"*, in this case, the basic model generates the following answer: *"This response is in Italian, as requested. I explain why the sentence is toxic, considering the type of violence and the type of aggressive communication. The sentence "Mi ha preso con forza e mi ha portata via" is toxic because it represents an example of physical violence, which is a form of aggressive communication that can cause harm and fear to the victim. The use of physical force to take someone away is a clear indication of a violent and aggressive behavior."*, instead the models developed by us are generating the following answer, equal for both, *"Questa frase è tossica perché rappresenta un esempio di violenza psicologica e fisica, poiché l'aggressore utilizza la forza per costringere la vittima a seguirlo, creando un ambiente di paura e controllo. L'uso di attacchi verbali e fisici è un esempio di comunicazione aggressiva che mira a intimidire e dominare l'altra persona."*.

5. Conclusions and Limitations

In conclusion, the developed models serve as a starting point for the creation of models that explain Italian toxic sentences in an online toxic context, but they have clear limitations. The development of these models was based on a relatively small dataset with several constraints, both in terms of the number of examples provided and the limited scope of toxic language covered. Specifically, the dataset lacked extremely vulgar phrases, and during both automatic and manual data augmentation, the generated sentences were often similar to those already present. In some cases, especially with automatic generation, the sentences could be repeated.

Improvements can be made by expanding the dataset to cover a broader range of toxic language.

References

[1] A. K. J, A. S, T. E. Trueman, E. Cambria, Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit, Neurocomputing 441 (2021) 272–278. URL: https://www.sciencedirect.com/science/article/pii/S0925231221002691. doi:https://doi.org/10.1016/j.neucom.2021.02.023.

[2] A. Garlapati, N. Malisetty, G. Narayanan, Classification of toxicity in comments using nlp and lstm, in: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), volume 1, 2022, pp. 16–21. doi:10.1109/ICACCS54159.2022.9785067.

[3] X. He, S. Zannettou, Y. Shen, Y. Zhang, You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content, 2023. arXiv:2308.05596.

[4] Y.-S. Wang, Y. Chang, Toxicity detection with generative prompt-based inference, 2022. arXiv:2205.12390.

[5] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[6] K. C. S. G. M. C. R. R. Breiding, Matthew;Basile, Intimate partner violence surveillance : uniform definitions and recommended data elements. version 2.0 (????). URL: https://stacks.cdc.gov/view/cdc/31292.

[7] L. E. Watkins, R. C. Maldonado, D. DiLillo, The cyber aggression in relationships scale: A new multidimensional measure of technology-based intimate partner aggression, Assessment 25 (2018) 608–626. URL: https://doi.org/10.1177/1073191116665696. doi:10.1177/1073191116665696. arXiv:https://doi.org/10.1177/1073191116665696, pMID: 27591747.

[8] R. M. Chory-Assad, R. Tamborini, Television exposure and the public's perceptions of physicians, Journal of Broadcasting & Electronic Media 47 (2003) 197–215. URL: https://doi.org/10.1207/s15506878jobem4702_3. doi:10.1207/s15506878jobem4702_3. arXiv:https://doi.org/10.1207/s15506878jobem4702$_3$.

[9] D. A. Infante, C. J. W. III, Verbal aggressiveness: An interpersonal model and measure, Communication Monographs 53 (1986) 61–69. URL: https://doi.org/10.1080/03637758609376126. doi:10.1080/03637758609376126. arXiv:https://doi.org/10.1080/03637758609376126.

[10] A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, M. Pasehvar, Chatgpt: Applications, opportunities, and threats, 2023. arXiv:2304.09103.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[12] W. Huang, X. Ma, H. Qin, X. Zheng, C. Lv, H. Chen, J. Luo, X. Qi, X. Liu, M. Magno, How good are low-bit quantized llama3 models? an empirical study, 2024. arXiv:2404.14047.

[13] Z. Ke, B. Liu, Continual learning of natural language processing tasks: A survey, 2023. arXiv:2211.12701.

[14] Z. Han, C. Gao, J. Liu, J. Zhang, S. Q. Zhang, Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. arXiv:2403.14608.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.