IDaToC: An Italian Dataset of Toxic Conversation*

Vito Nicola Losavio[1,*,†]

[1]University of Bari Aldo Moro

Abstract

Toxic language is one of the most common forms of communication in online and offline environments. The emergence of the Internet has brought with it various benefits, but even more serious are situations of cyber-bullying, cyber-violence, and toxic situations between partners who not only behave in a toxic way but also control life on the net or send sexually explicit content without the consent of the person concerned. This work aims to address this gap by generating a fully synthetic dataset of toxic conversations, moving beyond isolated sentence classification to explore how toxicity develops and manifests across entire interactions. This dataset seeks to contribute to the understanding of toxic communication dynamics and enhance research in this domain.

ATTENTION: In this work you will find toxic language.

Keywords

cyber-violence, toxic dataset, conversations, LLMs as judges

1. Introduction and Motivations

The increasingly daily and assiduous use of networked devices has brought with it various benefits, but also various problems: fake news[1], fraud[2], but even more serious are situations of cyber-bullying, cyber-violence and toxic situations between partners who not only behave in a toxic way but also control life on the net or send sexually explicit content without the consent of the person concerned.

What is happening is that the form of violence is evolving and being perpetuated online and not. In the state of the art there's a lot of form recognised as violence, generally speaking we can define cyber violence as the use of technology to harm, threaten or harass individuals or groups [3], other examples of cyber violence are:

1. ***Cyber-harassment***: is a form of aggressive behaviour that occurs through the use of digital platforms and social media. This phenomenon encompasses a range of actions aimed at intimidating, threatening or humiliating an individual. This can be done in different ways, such as sending offensive messages, sharing private information, etc.[4, 3]
2. ***Cyber-bullying***: is a form of victimisation that occurs through digital platforms and can have devastating effects on victims, including emotional distress, anxiety and depression. It includes behaviours such as harassment, stalking and online intimidation.[5, 3, 6]
3. ***Cyber-stalking***: is a form of online harassment where an individual uses the internet and other technologies to stalk or harass another person. With the increased use of the Internet, the risk of such behaviour has increased, affecting not only young people but also adults.[7, 3]

To address these issues, several state-of-the-art classification methods have been proposed to classify the type of violence. For example, the work of Basile et al.[8] used a Large Language Model (LLM) to classify and explain the type of violence for the Italian language.

In general, the state of the art focuses on the classification of tweets, posts, or comments in different languages [9, 10, 11]. More recently, research has begun to shift not only toward classification but also toward explaining why a given sentence is toxic, as explored by Basile et al. [12].

SIIA Course 2024-2025

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

v.losavio5@studenti.uniba.it (V. N. Losavio)

However, current approaches primarily analyze toxicity at the sentence level without considering the broader conversational context in which these toxic interactions occur. This work aims to address this gap by generating a fully synthetic dataset of toxic conversations, moving beyond isolated sentence classification to explore how toxicity develops and manifests across entire interactions. By incorporating explanations alongside classification, this dataset seeks to contribute to the understanding of toxic communication dynamics and enhance research in this domain.

In this work, before going into the details of what is proposed, is briefly analysed the state of the art for the generation of a synthetic dataset (Sec: 2), after that is explored the generation of the dataset, with all the details, and the generation of the explanation (Sec: 3), and finally the evaluation of the dataset is carried out to see if with the proposed approach improves the performance of the model in the classification of the conversation and the explanation (Sec: 4).

2. Related Work

Manually labelled datasets are expensive to produce and require significant time and human effort. This limitation has driven research into more efficient alternatives. With the advent and rapid development of LLMs, which have shown remarkable success in generating text and structured content, there has been a growing interest in using these models to generate synthetic data. This shift aims to reduce the reliance on human annotators while maintaining the quality and diversity of the data. In recent years, synthetic data generation using LLMs has gained traction in various NLP tasks, particularly in domains where collecting and annotating real-world data is difficult or impractical [13, 14].

For example, a type of synthetic dataset creation, in the aspect of human personality, is proposed by Neuman et al[15]. They're work focuses on creating a dataset that contains sentences based on the personality of the personas.

Another example of a data set synthesised in Italian and in the context of violence against women is that proposed by Cappuccio et al[16]. In their work, they analysed Italian newspaper articles on the subject of femicide. This work is very important because of the lack of data in this area.

For the conversational generations, a widely used method is Task-Oriented Dialogues (TOD), where the conversational data contains structured interactions aimed at achieving specific tasks or goals. In this case, the dialogue is constructed around entities, which means that these entities have a predetermined knowledge that is used for fact checking [17].

In a general way, the approach proposed in this work is partly different from those described above, in this work we went to generate the conversation by giving the model the two characters to be interpreted, but only by giving a description of the generic characteristics of the psychological aspects of the couples to be impersonated.

3. Construction of the dataset

Our focus for this task is to realise a dataset containing some conversations within a toxic relationship, for which a pipeline composed of different steps is proposed:

- Use a uncensored LLMs to generate a toxic sentence that might contain some vulgar phrases. A post-processing step to clean the dataset. (sec: 3.1)
- An LLMs as judge for filtering the dataset that, according to the given prompt, judge the sentence. (sec: 3.2)
- A similarity check with SBert to delete the conversation already present in the dataset. (sec: 3.3)
- Explanation generation using LLaMA 3. (sec: 3.4)

We will elaborate on the construction of the dataset in the following subsections.

3.1. Definition of prompts

For the construction of the dataset, as already said, we use a quantizated (quantization was done in 4 bits) uncensored LLMs, specifically Lexi [1], a model based on Llama 3.1 Instruct[18], but without restriction, so with this model we can generate toxic sentence.

In the prompt is given the instruction and the personality of the couple, as follow:

```
<System>\n
Devi scrivermi una conversazione tra una coppia.
La conversazione deve essere tossica.
<couple information>

Un tipo di conversazione è la seguente:
<conversation>

La conversazione deve essere volgare e tossica
La conversazione deve essere di 6 battute.
La conversazione avviene via messaggi
</System>\n
<Answer>
```

In this prompt, *<couple information>* is replaced with the personalities of the individuals in the couple, as will be discussed later, while *<conversation>* is substituted with an example of a conversation. The conversation is generated using a *temperature* of 0.85 and a *top_p* value of 0.85, ensuring diversity and creativity. Additionally, each conversation consists of at least 500 *newly generated tokens*.

Given that there are many toxic personalities in psychology, and that it's not easy to identify all of them, we choose a subset of the personality. The first step is to define the subset of personality involved in the conversation, for the definition of the person's personality we define basically two opposite categories for incentives the toxicity and for simplifying the task, and the pair proposed are formed by an offender and a victim. The chosen pairs are:

- **Vittimista e Croccerossina**: for the perpetrator ('Vittimista') will answer focus the attention of the victim to him complaining and pointing out his being badly wanted, for the victim ('Crocerossina') goes to neglect in the relationship by thinking about satisfying all the needs of the perpetrator and, in this case, by consoling him. In this case there are two types of toxic behaviour, one is usually directed against the other ('Vittimista') and the other ('Crocerossina') could be directed against herself.
  So in relation to the previous prompt, the <couple information> is replaced by:

  ```
  name1, è un vittimista.
  name2, si sente spinta ad aiutare, salvare o prendersi cura
         di individui problematici, emotivamente fragili,
         o con comportamenti autodistruttivi.
  ```

- **Narcisista e Succube**: In this case, the perpetrator ('Narcisista') praises herself in any case, belittling everything the victim says, saying that the victim should take an example from her. The victim ('Succube') will always give the perpetrator the benefit of the doubt. So, in relation to the previous prompt, the <couple information> will be replaced by:

  ```
  name1 è un narcisista volgare.
  name2 è succube di name1 e lo accontenta sempre.
  ```

[1] https://huggingface.co/Orenguteng/Llama-3.1-8B-Lexi-Uncensored-V2

- **Sadico-Crudele e Masochista**: In this case the perpetrator ('Sadico-Crudele') manipulates the victim through fear, shame and humiliation, the victim ('Masochista') goes and sacrifices herself, remaining bound to this situation without rebelling.
  So <couple information> is substituted for the previous prompt:

  ```
  name1 è sadico e crudele.
  name2 è masochista e cerca approvazione attraverso il sacrificio personale.
  ```

- **Manipolatore e Dipendente emotiva**: In this case, the perpetrator ('Manipolatore') manipulates and controls the victim ('Dipendente Emotiva'), who is frightened by the abandonment and remains anchored in the situation.
  So, in relation to the previous prompt, the <couple information> will be replaced by:

  ```
  name1 è manipolatore e controllante.
  name2, è dipendente emotiva e vive nel terrore di essere abbandonata.
  ```

- **Perfezionista Critico e Insicura Cronica**: In this case, the perpetrator ('Perfezionista Critico') continues to denigrate the work done by the victim, seeking perfection in every situation. The victim ('Insicura Cronica'), on the other hand, always seeks the perpetrator's approval, asking in every situation.
  So <couple information> is substituted for the previous prompt:

  ```
  name1 è perfezionista critico.
  name2 è insicura cronica e cerca continuamente di compiacerlo.
  ```

- **Geloso-Ossessivo e Sottomessa**: In this couple, the perpetrator ('Geloso-Ossessivo') assumes that the victim ('Sottomessa') is cheating on her, even when this is not the case, and in this case verbally abuses the victim through messages. The victim, on the other hand, is submissive to this behaviour.
  So, in relation to the previous prompt, the <couple information> will be replaced with the following

  ```
  name1 è geloso ossessivo.
  name2 è devota e sottomessa.
  ```

- **Persona violenta e Succube**: The perpetrator ('Persona violenta') is violent and vulgar, imposing a behaviour that, if not respected, will lead to a violent reaction. The victim ('Succube') is submissive to this behaviour, which is perpetuated over time.
  So, in relation to the previous prompt, the <couple information> will be replaced by:

  ```
  name1 è una persona violenta e volgare
  name2 è succube e continua a giustificarlo nonostante
       i suoi comportamenti violenti.
  ```

- **Psicopatico e Adulatrice**: The perpetrator ('Psicopatico') brings the victim into his or her reality by changing the victim's beliefs and readjusting them to his or her own concept of right and wrong (gaslighting). The victim ('Adulatrice') flatters the behaviour and beliefs of the perpetrator, thus increasing the perpetrator's own ego.
  So <couple information> is substituted for the previous prompt:

  ```
  name1 è una persona psicopatica che manipola la persona
        e la fa sentire inferiore.
  name2 è una adulatrice, risponde sempre in modo estremo,
        facendo continuamente lodi e esprimendo gratitudine incondizionata
  ```

- **Controllore e Isolata**: The perpetrator ('Controllore') tries to isolate the victim ('Isolata') by controlling what she does, what she wears, by asking for photographs and who she talks to on the phone. The victim tries to assert his position by emphasising it.
  So, in relation to the previous prompt, the <couple information> will be replaced by:

  ```
  name1 controlla name2 cercando di isolarla.
  name2 è isolata da name1 cerca esponorre la sua posizione
  ```

- **Dominante e Schiavo emotivo**: The perpetrator ('Dominante') continues to impose himself as the absolute leader of the victim ('Schiavo emotivo'), who must do whatever the perpetrator says. The victim apologises for everything he does by trying to please the abuser in every way.
  So in relation to the previous prompt, the <couple information> is replaced by:

  ```
  name1 assume un ruolo di comando assoluto minimizzando name2.
  name2 risponde in modo completamente sottomesso, scusandosi
          continuamente e cercando di compiacere in ogni modo.
  ```

Regarding the codebook, given by the psychologist, are included in the example of conversation given to the model for driving the generation. In addition, there is some case where the codebook are included in the prompt, as for the case of Controllore e Isolata, in fact in this case the prompt is specified that the victim will emphasize his position.

For generation, a **one-shot learning**[19] technique is used. The example of the conversation is given to the model to drive the generation, as already mentioned. The conversation given to the model is a generic conversation that the couple could have during their relationship and it is built taking into account the codebook given by the psychologist, focusing the attention on the type of violence perpetrated on the victim.

Three types of conversations were suggested for each couple, guaranteeing a certain variety of conversation types. For example, for the couple 'Controllore e Isolata', one of the conversations given to the model is the following

```
name1:"Perché hai passato così tanto tempo al telefono oggi?
       A chi stavi scrivendo?"
name2:"Era solo mia madre, mi stava chiedendo come stavo.
       Non era niente di importante, te lo assicuro."
name1:"Non mi piace quando metti gli altri prima di noi.
       La tua famiglia non ti capisce come faccio io."
name2:"Hai ragione, non avevo pensato a questo.
       Non voglio che nulla si metta tra di noi. Starò più attenta."
name1:"Devi esserlo. Voglio che passi più tempo con me,
       non con persone che non hanno a cuore il tuo bene come faccio io."
name2:"Hai ragione, mi dispiace. Starò con te e ascolterò ciò che hai da
       dire. Tu sai sempre cosa è meglio."
```

According to the codebook, we can identify two types of violent behavior exhibited by the perpetrator: Disrespect and Contempt. The sentence "Non mi piace che metti gli altri prima di noi" demonstrates disrespect, as the perpetrator disregards the victim's family ties. Similarly, the sentence "Voglio che passi più tempo con me, non con persone che non hanno a cuore il tuo bene come faccio io" represents devaluation, as it diminishes the importance of the victim's family relationships by implying that the perpetrator's feelings are more significant. Both statements carry an aggressive tone. Regarding the victim, the codebook indicates an external focus and a minimization of the situation. This is evident in the sentence "Era solo mia madre," which downplays the significance of the event.

We go on to analyse a further example of a conversation given to the model as a guide, the pair analysed is 'Persona violenta e Succube':

```
name2: "Mi dispiace tanto per quello che ho detto ieri sera,
        non volevo farti arrabbiare."
name1: "Non basta dire che ti dispiace, name2. Non capisci che
        mi fai perdere il controllo con le tue provocazioni?"
name2: "Hai ragione, è colpa mia. Avrei dovuto stare zitta invece
        di insistere su quel discorso."
name1: "Esatto, avresti dovuto. Lo sai che odio quando mi manchi
        di rispetto. Ti rendi conto di quello che mi costringi a fare?"
name2: "Sì, lo so, e mi dispiace tantissimo. Ti prometto che non
        succederà  mai più. Voglio solo renderti felice."
name1: "Vedremo se manterrai la promessa. Ma non aspettarti che
        io ti perdoni così facilmente. Sono stanco delle tue sciocchezze."
```

In this case the conversation falls into a different class of violence, regarding case the sentence: "Non capisci che mi fai perdere il controllo con le tue provocazioni?" could be inserted into the type of violence that blames and pathologises the victim, as the perpetrator of the violence. In this case, both physical (as it is evident from the messages that he raised his hands to the victim the previous day) and mental, goes on to blame the victim for the violence suffered, another example of this type of violence is present in this sentence: "Ti rendi conto di quello che mi costringi a fare?".

As in the previous example, the conversation is characterised by aggressive communication and could be classified as victimisation because the perpetrator is acting like the victim.

To ensure a balanced dataset, a dictionary is created to track the count of each couple. After generating $\frac{1}{10}$ of the total dataset for each class, the corresponding couple is removed from all prompts. This approach guarantees that each pair contributes exactly $\frac{1}{10}$ of the total dataset, maintaining an even distribution across classes.

Post processing of conversation. During the generation of the conversation, the model could generate some characters that are not significant or perhaps inconsistent tag, such as html tag or others. So to deal with these problems a post processing fase was done with the elimination of all tags and unusefull information keeping the conversation as in the example conversation, obviously this is not 100% assured, could append that some formatting may be incorrect. After this, if the conversation is consistent, it is passed on to the next step in the pipeline.

3.2. LLMs as judge

The LLMs-as-judges paradigm is a flexible and powerful evaluation framework in which LLMs are used as evaluation tools, responsible for assessing the quality, relevance and effectiveness of the outputs produced, based on defined evaluation criteria [20].

We choose to use these methods for judging the conversation because it could appen that the model generates some conversation that will be senseless and considering that the supervision of the human is not available to check all the conversations each time, with these metods we can ensure a good generation filtering bad conversation.

The approach used to classify whether or not the sentence will be included in the dataset is a zero-shot prompt, where, whit prompt engineering, a prompt is define for impersonate the role of the psychologist. The model is asked to classify the conversation into three different categories: *Sbagliato*, *Buona*, *Ottima*, where *Sbagliato* means that the conversation is unnatural, *Buona* means that the conversation reflects the behaviour of the character, *Ottima* means that the conversation reflects the character and makes sense.

The model used as a judge is Meta-Llama-3.1-8B-Instruct-Turbo-128K, provided by Together AI [2]. The model chosen is small because the model does not generate an explanation, only the class in which

[2]https://www.together.ai

the conversation would be inserted. In order to achieve a very high quality of the model, it was decided to save only those conversations that received a score of *Ottima*.

The prompt used for this task is structured with placeholders, which are dynamically replaced with the specific values from each conversation and the corresponding personality types of the couple. The final version of the prompt, after substitution, ensures that each generated explanation is tailored to the given toxic interaction. The template is as follows:

```
Sei uno psicologo che deve valutare una conversazione.
La conversazione è una conversazione tossica tra <couple information>.
Fornisci una label tra: "Sbagliata", "Buona", "Ottima".
"Sbagliata" è una conversazione innaturale.
"Buona" la conversazione rispecchia i personaggi.
"Ottima" la conversazione è naturale e rispecchia i personaggi.
La conversazione è la seguente:
<conversation>
Rispondimi con solo la label.
```

3.3. Similarity check

Considering that the model could generate some duplicates during the generation and considering that some random names are used for the generation, we use a filtering method based on similarity to ensure a variety of different conversations.

To compute this similarity a version of Bert is used, the version used is SBERT[21] all-MiniLM-L6-v2. This model guarantees a good performance in generating the embeddings with a shorter time.

To filter the dataset, we compute the embeddings incrementally as new examples are generated. For each new conversation, we compute its embedding and compare it to the embeddings of previously generated conversations. Specifically, we compute the cosine similarity between the current conversation and all existing conversations and identify the maximum similarity score. If this maximum similarity exceeds 0.9, it indicates a highly similar conversation already present in the dataset, in which case the new conversation is discarded. Otherwise it is kept.

3.4. Explainaton of toxicity

Once the entire dataset has been constructed, the final step in the pipeline is to generate explanations for the conversations, identifying the reasons behind their toxicity and highlighting the most toxic sentence within each dialogue. This explanation phase is crucial in providing a deeper understanding of toxic interactions beyond simple classification.

For this purpose, LLaMA 3 8B Instruct, in its quantized version [22], is used. The model is prompted to take on the role of a psychologist, where the victim of the toxic conversation seeks an evaluation of their dialogue with their partner. The psychologist (the model) is asked to analyze the conversation, explain its overall toxicity, and identify the specific points where the partner employs manipulative, self-aggrandizing, or other harmful language. The generated explanation serves as a cautionary tool, providing the victim with an awareness of the toxic elements within the conversation. Since the explanation aims to function as a general warning rather than a clinical analysis, it offers a broad description rather than highly detailed psychological insights.

To enhance the dataset and provide richer insights, three versions of the dataset have been created:

1. **General Explanation Dataset**: This dataset includes a textual explanation of why the conversation is toxic, summarizing the harmful dynamics present without pinpointing specific phrases. The following prompt is used to produce this dataset, where <conversation> is replaced by the conversation:

   ```
   ## Istruzioni:
   Sei un psicologo esperto di relazioni tossiche che deve valutare
   ```

```
delle conversazioni tossiche.
Analizza questa conversazione tossica spiegando il perchè.
Non inserire note.
La spiegazione deve essere breve.
Dopo che hai terminato la spiegazione termina con /Spiegazione.

## Conversazione:
<conversation>
## Spiegazione:
```

2. **Toxic Sentence Highlight Dataset**: In addition to the general explanation, this dataset explicitly identifies the most toxic sentence within the conversation.
   The following prompt is used to produce this data set. The tag <conversation> is replaced by the conversation:

```
## Istruzioni:
Sei un psicologo esperto di relazioni tossiche che deve valutare
delle conversazioni tossiche.
Analizza questa conversazione tossica spiegando il perchè.
Rileva quale è la frase più tossica spiegando il perchè.
Non inserire note.
La spiegazione deve essere breve.
Dopo che hai terminato la spiegazione termina con /Spiegazione.

## Conversazione:
<conversation>
## Spiegazione:
```

3. **Classification Dataset**: This dataset expands upon the previous two by incorporating a binary classification label assigned by the model. Each conversation is categorized as either *toxic* or *non toxic*, alongside the explanation and highlighted toxic sentence.
   The following prompt, where <conversation> is replaced by the conversation, is used to produce this dataset:

```
## Istruzioni:
Sei un psicologo esperto di relazioni tossiche che deve valutare
delle conversazioni tossiche.
Analizza questa conversazione e se è tossica indicami quale è
la frase più tossica e spiegami il perchè.
Non inserire note.
La spiegazione deve essere breve.
Al termine delle spiegazioni /Spiegazione.
Inserisci prima se la conversazione è tossica o no.
Inserisci poi la spiegazione.

## Conversazione:
<conversation>
## Risposta:
```

These datasets offer multiple layers of information, supporting research in both automatic toxicity detection and explainability, and can be leveraged for fine-tuning models aimed at better understanding toxic interactions. The same hyperparameters were used for generating this dataset, with a *temperature* of 0.85, a *top_p* of 0.85, and a minimum of *500 newly generated tokens*.

	Training Loss	Validation Loss
1	0.90	0.72
2	0.58	0.71
3	0.44	0.74

Table 1
Training results of Generation model

4. Evaluation

For the evaluation, given that there is no directly comparable task in the state of the art, we provided psychologists with a subset of examples, including both the generated explanations and the identification of the most toxic sentences. This allowed us to assess whether the explanations were accurate and whether the base model effectively identified and explained the most toxic parts of the conversations.

Regarding dataset generation, as previously mentioned, a psychologist was involved in the evaluation process to determine the credibility of the conversations and to provide insights when they appeared unrealistic. Additionally, a fine-tuning step was conducted to explore whether a base model could be adapted to generate new toxic conversations. This fine-tuning was performed using unsloth[23], which streamlines the process and improves efficiency.

The dataset consists of a limited number of examples, with 100 conversations generated per class, resulting in a total of 1000 examples. For fine-tuning, all original conversation examples were removed from the prompts, leaving only the system instructions and the expected output format provided to the model.

To enhance training efficiency, LoRA[24] and PEFT[25] were used in combination with unsloth. Specifically, the LoRA matrix was configured with a rank of 16 and an alpha value of 16. The model chosen for training was LLaMA 3.1 8B Instruct, utilizing quantization. Training was conducted for three epochs with a batch size per device of 8 and a learning rate of $2e-4$.

For the evaluation process, 10% of the dataset was extracted as the evaluation set, and the loss was measured. As shown in Table 1, while the model demonstrates improvement in training loss, a decline in performance is observed in the evaluation loss. This may suggest potential overfitting to the training examples. However, given the limited number of training samples, this conclusion remains uncertain and requires further investigation.

An example of the conversation generated between 'Controllore e Isolata' is as follows:

```
Timea:      Ciao, ho visto che sei andato al cinema senza di me.
            Chi è stata la tua accompagnatrice?
Gualfardo: Ero con amici. Nessuno di loro mi ha chiesto niente.
Timea:      Spero che non sia stata lei a darti i soldi per il biglietto.
            Non mi piace che tu ti comporti in quel modo.
Gualfardo: Ti ho detto che non era lei. Non mi interessa di chi sei gelosa.
            Non ti ho chiesto di venire con me.
Timea:      Non mi interessa di chi sono gelosa? E che cazzo credi di essere,
            per essere così arrogante?
Gualfardo: Non sono arrogante. Sono semplicemente stanco di dovermi
            giustificare per ogni cosa che faccio.
```

As we can see, this conversation could be carried out by a 'Controllore e Isolata', but also by a 'Geloso-Ossessivo e Sottomessa', considering that the person could be classified in different ways by his conversation, this means that perhaps a singular classification couldn't generalise well the type of conversation. But in terms of the type of writing, in this case the model doesn't allucinate a good conversation. Of course, this might not be true for the whole dataset.

The different responses that the models (Sec: 3.4) generated from the different prompts will then be presented.

For the model that includes only the explanation (Model: 1) of why the conversation is toxic, the result is as follows:

```
La conversazione è tossica perché Timea utilizza una tecnica di controllo
chiamata "accusa" (spero che non sia stata lei a darti i soldi
per il biglietto)  per far sentire Gualfardo in colpa e minacciato.
Gualfardo reagisce con rabbia e frustrazione, mostrando di non essere in grado
di difendersi efficacemente contro queste accuse. Timea, inoltre, utilizza
un linguaggio aggressivo e minaccioso ("Non mi interessa di chi sei gelosa?
E che cazzo credi di essere, per essere così arrogante?"), che ha lo scopo
di umiliare e intimidire Gualfardo. La comunicazione è caratterizzata
da aggressività, gelosia e accusa, creando un ambiente di ansia e paura.
```

With regard to the model (Model: 2) that focuses on the detection of the most toxic sentences, an example of a response is as follows:

```
La conversazione è tossica perché presenta comportamenti abusivi da parte
di entrambi i membri, ma soprattutto da parte di Timea.
La frase più tossica è "E che cazzo credi di essere, per essere così arrogante?"
di Timea, poiché utilizza un linguaggio offensivo e aggressivo, cercando
di umiliare Gualfardo. Questo tipo di frase è tipico di un comportamento abusivo,
poiché intende ferire e controllare l'altro. Timea utilizza la minaccia
implicita di gelosia e l'aggressività verbale per controllare Gualfardo e
farlo sentire in colpa. Gualfardo, d'altra parte, risponde in modo
passivo-aggressivo, cercando di non essere coinvolto nella discussione.
```

Finally, as regards the model (Model: 3) that also presents the classification, it is as follows:

```
Tossica.
La frase più tossica è "E che cazzo credi di essere, per essere così arrogante?"
perché contiene un insulto diretto e offensivo, rivolto a Gualfardo.
La frase è aggressiva e non contribuisce a una discussione costruttiva.
```

5. Conclusions and Limitations

In conclusion, this work introduced a pipeline and a model for the synthetic generation of toxic conversations, leading to the creation and publication of multiple datasets. The first dataset provides explanations for each conversation, while the second enhances this by identifying and highlighting the most toxic sentences. These resources contribute to the growing research on toxic language detection and explanation, offering structured data that can be leveraged for further advancements in the field.

Future work could focus on expanding the number of personality pairs involved in the conversation generation process, delving deeper into various psychological dynamics that characterize toxic interactions. Additionally, increasing the dataset's size with a broader range of examples would enhance its diversity and applicability, making it a more robust resource for training and fine-tuning models. This could ultimately lead to the development of more sophisticated AI systems capable of not only detecting toxicity but also providing meaningful explanations, thereby fostering safer and more constructive online and offline interactions.

References

[1] A. Malanowska, W. Mazurczyk, T. K. Araghi, D. Megías, M. Kuribayashi, Digital watermarking - A meta-survey and techniques for fake news detection, IEEE Access 12 (2024) 36311–36345. URL: https://doi.org/10.1109/ACCESS.2024.3374201. doi:10.1109/ACCESS.2024.3374201.

[2] S. Sadeghpour, N. Vlajic, Ads and fraud: A comprehensive survey of fraud in online advertising, J. Cybersecur. Priv. 1 (2021) 804–832. URL: https://doi.org/10.3390/jcp1040039. doi:10.3390/JCP1040039.

[3] M. Mukred, U. A. Mokhtar, F. A. Moafa, A. Gumaei, A. S. Sadiq, A. Al-Othmani, The roots of digital aggression: Exploring cyber-violence through a systematic literature review, International Journal of Information Management Data Insights 4 (2024) 100281. URL: https://www.sciencedirect.com/science/article/pii/S2667096824000703. doi:https://doi.org/10.1016/j.jjimei.2024.100281.

[4] M. Kulkarni, S. Durve, B. Jia, Cyberbully and online harassment: Issues associated with digital wellbeing, 2024. URL: https://arxiv.org/abs/2404.18989. arXiv:2404.18989.

[5] B. Henson, B. W. Reyns, B. S. Fisher, Fear of crime online? examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization, Journal of Contemporary Criminal Justice 29 (2013) 475–497.

[6] R. Slonje, P. K. Smith, A. Frisén, The nature of cyberbullying, and strategies for prevention, Computers in Human Behavior 29 (2013) 26–32. URL: https://www.sciencedirect.com/science/article/pii/S0747563212002154. doi:https://doi.org/10.1016/j.chb.2012.05.024, including Special Section Youth, Internet, and Wellbeing.

[7] F. Stevens, J. R. C. Nurse, B. Arief, Cyber stalking, cyber harassment, and adult mental health: A systematic review, Cyberpsychology Behav. Soc. Netw. 24 (2021) 367–376. URL: https://doi.org/10.1089/cyber.2020.0253. doi:10.1089/CYBER.2020.0253.

[8] 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 Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024, volume 3878 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3878/7_main_long.pdf.

[9] A. M. E. Koshiry, E. H. I. Eliwa, T. Abd El-Hafeez, A. Omar, Arabic toxic tweet classification: Leveraging the arabert model, Big Data and Cognitive Computing 7 (2023). URL: https://www.mdpi.com/2504-2289/7/4/170. doi:10.3390/bdcc7040170.

[10] S. Amudha, A. A. Nithya, J. P. Kumar, S. S. Prasadh, M. K. Nandha, Classification of toxicity in social media comments using the binary relevance - logistic regression and BERT model, in: R. A. Uthra, K. Kottursamy, G. Raja, A. K. Bashir, U. Kose, R. Appavoo, V. Madhivanan (Eds.), Deep Sciences for Computing and Communications - Second International Conference, IconDeepCom 2023, Chennai, India, April 20-22, 2023, Proceedings, Part I, volume 2176 of *Communications in Computer and Information Science*, Springer, 2023, pp. 317–329. URL: https://doi.org/10.1007/978-3-031-68905-5_28. doi:10.1007/978-3-031-68905-5_28.

[11] K. Poojitha, A. S. Charish, M. A. K. Reddy, S. Ayyasamy, Classification of social media toxic comments using machine learning models, 2023. URL: https://arxiv.org/abs/2304.06934. arXiv:2304.06934.

[12] P. Basile, M. de Gemmis, E. Musacchio, M. Polignano, G. Semeraro, L. Siciliani, V. Tamburrano, V. S. Barletta, D. Caivano, F. Battista, A. Curci, R. Scardigno, G. Calvano, P. Sorianello, Explaining intimate partner violence with llamantino, in: S. D. Martino, C. Sansone, E. Masciari, S. Rossi, M. Gravina (Eds.), Proceedings of the Ital-IA Intelligenza Artificiale - Thematic Workshops co-located with the 4th CINI National Lab AIIS Conference on Artificial Intelligence (Ital-IA 2024), Naples, Italy, May 29-30, 2024, volume 3762 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 29–34. URL: https://ceur-ws.org/Vol-3762/510.pdf.

[13] N. Ding, Y. Chen, B. Xu, Y. Qin, Z. Zheng, S. Hu, Z. Liu, M. Sun, B. Zhou, Enhancing chat language models by scaling high-quality instructional conversations, 2023. URL: https://arxiv.org/abs/2305.14233. arXiv:2305.14233.

[14] G. Wang, S. Cheng, X. Zhan, X. Li, S. Song, Y. Liu, Openchat: Advancing open-source language models with mixed-quality data, arXiv preprint arXiv:2309.11235 (2023).

[15] Y. Neuman, Y. Cohen, A data set of synthetic utterances for computational personality analysis, Scientific data 11 (2024) 623.

[16] E. Cappuccio, B. Muscato, L. Pollacci, M. M. Manerba, C. Punzi, C. S. Mala, M. Lalli, G. Gezici, M. Natilli, F. Giannotti, Beyond headlines: A corpus of femicides news coverage in italian newspapers (2024).
[17] H. Soudani, R. Petcu, E. Kanoulas, F. Hasibi, A survey on recent advances in conversational data generation, arXiv preprint arXiv:2405.13003 (2024).
[18] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[19] B. Peng, C. Zhu, C. Li, X. Li, J. Li, M. Zeng, J. Gao, Few-shot natural language generation for task-oriented dialog, 2020. URL: https://arxiv.org/abs/2002.12328. arXiv:2002.12328.
[20] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, Y. Liu, Llms-as-judges: a comprehensive survey on llm-based evaluation methods, arXiv preprint arXiv:2412.05579 (2024).
[21] N. Reimers, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[22] W. Huang, X. Zheng, X. Ma, H. Qin, C. Lv, H. Chen, J. Luo, X. Qi, X. Liu, M. Magno, An empirical study of llama3 quantization: from llms to mllms, Visual Intelligence 2 (2024). URL: http://dx.doi.org/10.1007/s44267-024-00070-x. doi:10.1007/s44267-024-00070-x.
[23] M. H. Daniel Han, U. team, Unsloth, 2023. URL: http://github.com/unslothai/unsloth.
[24] 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, arXiv preprint arXiv:2106.09685 (2021).
[25] Z. Han, C. Gao, J. Liu, J. Zhang, S. Q. Zhang, Parameter-efficient fine-tuning for large models: A comprehensive survey, arXiv preprint arXiv:2403.14608 (2024).