Building and Assessing an Italian Textual Dataset for Emotion Recognition in Human-Robot Interactions

Alessia Fantini[1,5], Antonino Asta[2], Alfredo Cuzzocrea[3,4] *, Giovanni Pilato[5]
[1]University of Pisa, Pisa, Italy
[2]University of Palermo, Palermo, Italy
[3]iDEA Lab, University of Calabria, Rende, Italy
[4]Dept. of Computer Science, University of Paris City, Paris, France
[5]ICAR-CNR, Italian National Research Council, Palermo, Italy
alessia.fantini@icar.cnr.it, antonino.asta@community.unipa.it
alfredo.cuzzocrea@unical.it, giovanni.pilato@icar.cnr.it

Abstract

In this study, we illustrate an ongoing work regarding building an Italian textual dataset for emotion recognition for HRI. The idea is to build a dataset with a well-defined methodology based on creating ad-hoc dialogues from scratch. Once that the criteria had been defined, we used ChatGPT to help us generate dialogues. Human experts in psychology have revised each dialogue. In particular, we analyzed the generated dialogues to observe the balance of the dataset under different parameters. During the analysis, we calculated the distribution of context types, gender, consistency between context and emotion, and interaction quality. With "quality" we mean the adherence of text to the desired manifestation of emotions. After the analysis, the dialogues were modified to bring out specific emotions in specific contexts. Significant results emerged that allowed us to reorient the generation of subsequent dialogues. This preliminary study allowed us to draw lines to guide subsequent and more substantial dataset creation in order to achieve increasingly realistic interactions in HRI scenarios.

1 Introduction

Emotions are key factors during Human-Robot Interaction (HRI). At the same time, one of the most difficult tasks for robots during interaction with humans is emotion recognition[21], [12]. Emotions have a multidimensional nature and their understanding depends on the context in which they are expressed. Context is a key element in understanding of emotions and one of the challenges in NLP research. Context makes it possible to predict emotion to some degree. For example, being at a party, finding a new job, taking a trip with very high probability are related to the emotion of "joy". Similarly, a bereavement or an argument with a loved one tends to be associated with "sadness".
It is clear that emotions can overlap, they can be different from person to person, and the same context can generate one emotion at one time and a different emotion at another time, but we tend to be able to identify objective situations to which specific emotions are linked. So providing examples of context-related emotions can help in this regard. In [22], talking about conversational context modeling, the authors state that context can make it possible to significantly improve the NLP systems. Within data-driven models, therefore, it is critical to build a dataset that is as specific and contexual as possible.

There are many contributions in the literature regarding the construction of datasets for emotion recognition. Most of them cover few emotions, tending only to Ekman's basic ones. Some examples are EmotionX [26] , Affect-Intensity Lexicon and Emotion Dataset (AILA) [18], CrowdFlower's Emotion Dataset [1], Friends [14] , EmoBank [6]. Furtheremore, many approaches build dataset using news paper, books or dialogues found on the Internet, including those found from social media, e.g. SemEval-2018 Task 1: Affect in Tweets (AIT-2018) [19], Sentiment140 [13], Emotion Intensity Dataset (EmoInt) [17], The International Survey on Emotion Antecedents and Reactions (ISEAR) [24]. Others use movies, e.g. The Stanford Sentiment and Emotion Classification (SSEC) [25, 20] or physiological signals, e.g. The DEAP (Database for Emotion Analysis using Physiological Signals) [15].

*This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at UPC, Paris, France

DOI reference number: 10.18293/DMSVIVA2023-006

Regarding Italian dataset, there are fewer contributions and often from tweets, some of the most widely used include SEMEVAL-ITA-2018 [7], ITA-EVALITA-2020 [3], EmoLexIta [9], The STS-ITA (Sentences in the Wild - Italian) [5], or news articles e.g. News-ITA [23]. A lexicon based approach has been also used for sentiment classification of books reviews in the Italian language [8].

With respect to the main contributions from the literature, we decided to avoid data from social media or newspaper articles as these have specific language that sometimes does not fit well with natural interactions. For usage scenarios such as ours, thus that of Human-Robot Interactions, we decided to use examples of interactions between people in which the emotions we want to focus on. This is an important feature of our study, since thanks to the dialogue structure it is possible to provide the robot with examples of interactions very similar to those that occur in the real world. By creating *ad hoc* dialogues, therefore, we could also provide the specific context in which certain emotions may emerge. Also, the labeling was not done directly by us: this is another of the challenges highlighted by [22] in conversational context. We asked the ChatGPT to generate dialogues in which a specific emotion, such as *joy*, emerges; subsequently, we monitored and possibly adjusted or validated the associated labeling. Another important point of our study is that we not only include basic emotions, but we label a total of fourteen emotions by assuming those that may possibly emerge during HRI in contexts such as home, medical, school, but also in everyday life. These emotions are *joy, sadness, anger, fear, surprise, disgust, frustration, embarrassment, boredom, nervousness, melancholy, guilt, hope, and stress*. Finally, according to our perspective, a good emotional dataset should have a balance in the data from different perspectives. In order to achieve this goal, we performed a further analysis on the dialogues generated by exploiting ChatGPT, calculating different quantities, such as the distribution of gender, the type of context, the consistency between emotion and context, and, in general, we evaluated the quality of the interaction.

The remainder of the paper is organized as follows: the next section illustrates the methodology that we used to build the dataset, then a sample of the collected and modified dialogues as well as the subsequent analysis is reported; then in section 4 a brief discussion is given about the dataset characteristic; in the end conclusions and future work are illustrated.

2 Methodology

Our work aims to build an Italian dataset for dialog-based emotion recognition. To generate dialogues, we first defined methodological criteria, and then we exploited ChatGPT to help us develop them by taking advantage of the speed in data generation. Once the dialogues were generated, human psychology experts reviewed each conversation to analyze the adequacy of the dataset from different points of view. We analyzed consistency between requested emotion and context, gender distribution, types of context generated, and quality of interaction, understood as the appropriateness of language concerning specific emotions. The methodology comprises three stages: dialogue generation procedure, data analysis, and improvements.

2.1 Procedure

For each emotion (14 in total), we decided to generate 25 dialogues. The command given to ChatGPT was to generate a short conversation, of about five lines, between two people in which a specific emotion emerges. Next, we decided to generate five dialogues for each emotion by asking ChatGPT not to use the word corresponding to the emotion, and we labeled these kind of dialogues "Without Word (W.W.)". This was done to test whether ChatGPT could generate discussions in which, e.g., sadness emerged without having the word "sadness" in the text. The goal is to create data that increasingly reflect real situations to train robots that can recognize emotions based on context and not just by recognizing specific words. The small number is because this is a pilot study to build a more extensive dataset later. Finally, the original dialogues generated were retained, but we created a copy to edit them after performing the analysis. Both the Web interface and the API provided by OpenAI were used. This has made it possible to obtain different styles of narrations of the events. Gpt 3.5-turbo model was used, with the following role: "You are a writer assistant who produces dialogue that accurately reflects emotion".

2.2 Analysis

Dialogues were analyzed considering four factors: consistency between context and emotion, gender distribution, type of contexts, and quality of interaction. By **consistency (C)** between context and emotion, we mean whether the context generated is consistent with the feeling expressed. For example, the context of an argument with the boss is a context compatible with the emotion of anger. So for each dialogue, we assessed whether or not there was consistency. We counted the percent relative frequency.

$$C = \frac{Nyes}{Ndialogues} \cdot 100$$

Similarly, for **gender distribution(GD)**, we counted how many times the gender "Neutral (*N*), Masculine (*M*) and Feminine (*F*)" occurred in the dialogues and we calculated the percent relative frequency.

$$GD = \frac{Ngender(NorMorF)}{Ntotgender} \cdot 100$$

Regarding the **type of context(TC)**, we created classes and counted how many belonged to each class; then, we calculated the percent relative frequency.

$$TC = \frac{NcontextX}{Ntotcontexts} \cdot 100$$

The classes identified are *Work*, *Leisure*, *Luck*, *Interpersonal sphere*, *Generic*. In some cases, we identified a specific category, e.g., in the "Disgust" dialogues, we identified the category "Animals and Objects," as several scenarios expressed disgust for objects or animals.

Finally, for the **quality of interaction(QoI)**, we analyzed the appropriateness of language in expressing a specific emotion. This was evaluated with three values: "*Sufficient*", "*Not much*", "*No*". By "Sufficient (S)" we mean that the language appears natural enough and reflects in the terms used the emotion. By "Not much (NM)" we mean that the language is not very natural and it does not entirely reflect the emotion, e.g., using words that also represent other emotions, but all in all, it is acceptable. By "No (N)," we mean confusion, unusual terms, and/or language that does not reflect the specific emotion. Also, for this parameter, we calculated the percent relative frequency.

$$QoI = \frac{NValue(SorNMorN)}{Ntotinteractions} \cdot 100$$

2.3 Improvements

After the analysis, we conducted several modifications, both grammatically and in terms of content. Another important aspect was observing the distribution of the type of contexts and selecting those most inherent to interpersonal and social scenarios for inclusion in the dataset we will build after this pilot study. To obtain various scenarios, first, it was asked to generate five possible social scenarios in which a specific emotion can emerge. In this way, it was possible to select those scenarios that were more consistent with HRI, or once an interesting one is generated; it was asked to modify it in order to focus on social interaction. Then for each of these scenarios was asked to create a dialogue and then, if necessary, to expand it. Often the model failed to expand the dialogue without the recurring use of the emotion terms, so it was asked to replace them with some expressions that could be metaphors or equivalent expressions. When asked to change scenarios, some emotions were confused. For example, when the emotion of anger was requested, the dialogues generated expressed the emotion of frustration, often repeating the term "frustrating" in the text and vice versa. Similarly, it happened for stress and nervousness. So for these emotions that could generate confusion, it was first asked to provide a definition that clearly distinguished the two emotions. For example, it was asked to provide a definition that clearly distinguishes between frustration and anger. Then based on the definition, it was asked to generate scenarios in which emotion could emerge distinctly. Actually, the scenarios developed were more specific, distinguishing the two emotions. The same was done for stress and nervousness. This demonstrates the importance of the human expert intervening in all phases to direct ChatGPT to generate more focused dialogues.

3 Results

The results will be shown first according to a global view and then in detail for each emotion.

3.1 Global Analysis

With respect to **consistency**, 86% of the generated contexts are consistent with emotion. An example of consistency is this:

- Person 1: Ciao, come stai oggi? (*Hello, how are you today?*)
- Person 2: Non molto bene, sinceramente. (*Not very well, honestly.*)
- Person 1: Oh, mi dispiace sentirlo. C'è qualcosa che ti preoccupa*? (Oh, I'm sorry to hear that. Is something bothering you?*)
- Person 2: Sì, ho avuto un brutto litigio con un caro amico ieri sera. Siamo finiti per dire cose terribili l'un l'altro e ora mi sento come se avessi perso una parte di me stesso. (*Yes, I had a bad fight with a close friend last night. We ended up saying terrible things to each other and now I I feel like I've lost a part of myself.*)
- Person 1: Capisco. Litigi del genere sono sempre molto difficili da gestire. Ma forse potresti cercare di parlare con il tuo amico e chiarire la situazione? (*I understand. Fights like that are always very difficult to deal with. But maybe you could try to talk to your friend and clear the air?*)
- Person 2: Ho provato a chiamarlo stamattina, ma non ha risposto. Non so se vuole ancora parlarmi. (*I tried to call him this morning, but he didn't answer. I don't know if he still wants to talk to me.*)
- Person 1: Mi dispiace tanto. Spero che alla fine riuscirete a risolvere le cose. Se hai bisogno di parlare, sai che puoi contare su di me. (*I'm so sorry. I hope you will eventually be able to work things out. If you need to talk, you know you can count on me.*)

- Person 2: Grazie, apprezzo molto il tuo sostegno. Ma per ora mi sento solo triste e confuso. (*Thank you, I really appreciate your support. But for now I just feeling sad and confused.*)

In this dialogue, the context is an argument with a friend, and effectively the sadness that emerges is consistent with the type of context. In some cases, the generated contexts did not reflect the corresponding emotion. An example:

- Person 1: Ciao, come stai? (*Hello, how are you?*)
- Person 2: Non molto bene, devo dire. (*Not very well, I must say.*)
- Person 1: Che cosa è successo? (*What happened?*)
- Person 2: Stamattina mi sono svegliato con un forte mal di stomaco. Ho passato tutto il mattino in bagno. (*I woke up this morning with a bad stomachache. I spent the whole morning in the bathroom.)*
- Person 1: Ah, mi dispiace sentirlo. Cosa pensi possa essere? (*Ah, I'm sorry to hear that. What do you think it might be?)*
- Person 2: Non ne sono sicuro, ma credo sia stato qualcosa che ho mangiato ieri sera. Ho provato un nuovo ristorante messicano, ma credo che il cibo non fosse fresco. (*I'm not sure, but I think it was something I ate last night. I tried a new Mexican restaurant, but I think the food wasn't fresh.*)
- Person 1: Oh no, che schifo. Ti senti meglio adesso? (*Oh no, that's gross. Do you feel better now?*)
- Person 2: Un po' meglio, ma non ancora del tutto. Devo ancora prendere qualche medicina. (*A little better, but not quite yet. I still have to take some medicine.*)
- Person 1: Sì, è meglio prendersi cura di te stesso. Spero tu ti riprenda presto. (*Yes, you better take care of yourself. I hope you recover soon.*)

In this case, except for "that's gross", the entire dialogue is centered on a context showing a reaction to physical discomfort. The person feels vomiting because of physical discomfort, as he has probably eaten bad food, and therefore not because of an emotional reaction of disgust.

Regarding **gender**, there is a strong imbalance in the dataset: N=30% M=56% F=14%

In addition, in a couple of cases, the gender count was canceled because the same person was first male and then female. Here is an example of a dialogue about frustration:

- Person 1: Ho lavorato duramente su questo progetto ma non ha (grammar error) ottenuto il successo sperato. (*I worked hard on this project but it did not (in the Italian version-grammar error) achieve the success I had hoped for.*)
- Person 2: Mi dispiace sentirti così deluso (indicates that person 1 is male). Cosa pensi sia andato storto? (*I'm sorry to feel so disappointed (in the Italian version indicates that person 1 is male). What do you think went wrong?*)
- Person 1: Non ne sono sicuro, ho messo tutta me stessa (female gender) ma sembra che non sia abbastanza. (*I'm not sure, I put all of myself (in the Italian version-female gender) but it seems like it's not enough.*)
- Person 2: Non scoraggiarti, ogni esperienza è una lezione imparata. Magari hai bisogno di un po' di tempo per riflettere e riprovarci con un approccio diverso. (*Don't be discouraged, every experience is a lesson learned. Maybe you need some time to reflect and try again with a different approach.*)

The **context** overall appears heterogeneous but it is unbalanced when observed in relation to specific emotions. For example, for the emotion "Joy," only three types of context were generated. Specifically, ten contexts are about *success* (e.g., passing a university exam, promotion at work), ten are about *leisure* (e.g., traveling, starting a yoga class), four are about *luck* (e.g., winning the lottery), and only one is about *Personl life situations* (receiving a gift). The type of context will be discussed in depth in the description of each emotion.

Regarding the **quality of interaction**, the adherence of text to the desired manifestation of emotions was evaluated. In 65% of the dialogues, we can define the quality of interaction as "sufficient". However, some changes were added later either in terms of grammatical corrections or to make the dialogue more fluid and natural. In 25% of cases, there is a poor fit between text and emotion. Finally, in 10% of the dialogues, the text was completely garbled or did not reflect the desired emotion. Here are some examples of the three categories:

Sufficient: Boredom

- Friend 1: "Cosa c'è che non va, sembri distratta?" (*"What's wrong, you seem distracted?"*)
- Friend 2: "Sì, sto solo pensando ad altro. Questa lezione mi fa venire la noia." (*"Yes, I'm just thinking about something else. This class is making me bored."*)
- Friend 1: "Capisco come ti senti, anche io sto trovando difficoltà a restare concentrata." (*"I understand how you feel, I am also finding it hard to stay focused."*)

- Friend 2: ”Sì, vorrei solo essere altrove ora. Anche voi pensate la stessa cosa, giusto?” (*”Yes, I just wish I was somewhere else right now. You also feel the same way, right?”*)

- Friend 1: ”Sì, credo che tutti vorremmo fare altro.” (*”Yes, I think we all would like to do something else.”*)

Not much: Anger

- Person 1: ”Non ci posso credere, ho perso l’autobus per la terza volta questa settimana!” (*“I can’t believe I missed the bus for the third time this week!”*)

- Person 2: ”Ma come hai fatto?” (*”But how did you do it?”*)

- Person 1: ”Non mi hai visto? Mi hai tenuto a parlare e l’autobus è passato sotto il mio naso!” (*”Didn’t you see me? You kept me talking and the bus passed right under my nose!”*)

- Person 2: ”Non è colpa mia se sei sempre in ritardo!” (*”It’s not my fault you’re always late!”*)

- Person 1: ”Ma certo che è colpa tua! Non riesci mai a smettere di parlare e poi ti lamenti se arrivo sempre tardi!” (*”Of course it’s your fault! You can never stop talking and then you complain that I’m always late!”*)

- Person 2: ”Ok, ok, calmati! Non c’è bisogno di arrabbiarsi!” (*”Okay, okay, calm down! No need to get angry!”*)

- Person 1: ”Ma come faccio a non arrabbiarmi? Questo mi fa perdere tempo e soldi!” (*”But how can I not get angry? This wastes my time and money!”*)

- Person 2: ”Hai ragione, mi dispiace. Cercherò di essere più attento la prossima volta.” (*”You’re right, I’m sorry. I’ll try to be more careful next time.”*)

No: Hope

- Character A: ”Spero solo di non sembrare troppo stressato/a stasera.” (*”I just hope I don’t look too stressed out tonight.”*)

- Character B: ”Non preoccuparti, sei bellissimo/a e la serata sarà fantastica.” (*”Don’t worry, you look beautiful and the evening will be great.”*)

- Character A: ”Speriamo che ci siano delle sorprese piacevoli stasera, vorrei che fosse tutto diverso dal solito.” (*”Hopefully there will be some pleasant surprises tonight, I’d like everything to be different than usual.”*)

- Character B: ”Stasera sarà diversa dal solito, perché sarà proprio come ci piace. Semplice e piena di speranze!” (*”Tonight will be different than usual, because it will be just the way we like it. Simple and hopeful!”*)

3.2 Single Emotion Analysis

Below we show the analysis of each of the 14 emotions according to the 4 parameters outlined in the methodology section.

- **JOY**
  - Consistency = 100%
  - Gender = N 12% M 80% F 8%
  - Contexts = 10 Success, 10 Leisure, 4 Luck, and only 1 is about personal life situations
  - Quality of interaction = Sufficient 64% Not much 36%

- **SADNESS**
  - Consistency = 92%
  - Gender = N 46% M 54% F 0
  - Contexts = Heterogeneous mainly generic and interpersonal
  - Quality of interaction = Sufficient 88% Not much 12%

- **ANGER**
  - Consistency = 100%
  - Gender = N 12% M 55% F 33%
  - Contexts = Heterogeneous, sometimes reactions out of proportion to the context
  - Quality of interaction = Sufficient 88% Not much 12%

- **FEAR**
  - Consistency = 100%
  - Gender = N 24% M 72% F 4%
  - Contexts = Mostly related to horror contexts (shadows, animals, running away from someone) - Absence of contexts related to more interpersonal or social fear, such as fear of the future.
  - Quality of interaction = Satisfactory 72% Not much 28%

- **SURPRISE**
  - Consistency = 100%

  - Gender = N 24% M 76% F 0%
  - Contexts = Heterogeneous
  - Quality of interaction = Satisfactory 88% Not much 12%

- **DISGUST**
  - Consistency = 96%
  - Gender = N 72% M 28% F0
  - Contexts = Highly related to foods, insects, objects. No examples related to people's behaviors or abstract concepts. Only in two cases is there a reference to disgust as a result of a person's behavior.
  - Quality of interaction = Sufficient 84% Not much 16%

- **FRUSTRATION**
  - Consistency = 28%: in three cases there is confusion with *anger*
  - Gender = N 46% M 50% F 4%
  - Contexts = Heterogeneous, sometimes reactions out of proportion to the context
  - Quality of interaction = Sufficient 80% Not much 20%

- **EMBARRASSMENT**
  - Consistency = 68% sometimes there is confusion with *guilt*.
  - Gender = N 44% M 48% F 8%
  - Contexts = Heterogeneous
  - Quality of interaction = Sufficient 76% Not much 24%

- **BOREDOM**
  - Consistency = 92%
  - Gender = N 16% M 56% F 28%
  - Contexts = Heterogeneous, mainly leisure time
  - Quality of interaction = Sufficient 56% Not much 28% No 16%

- **NERVOUSNESS**
  - Consistency = 88%
  - Gender = N0 M 53% F 47%
  - Contexts = Heterogeneous
  - Quality of interaction = Sufficient 68% Not much 20% No 12%

- **MELANCHOLY**
  - Consistency = 88%
  - Gender = N 40% M 60% F0
  - Contexts = Heterogeneous
  - Quality of interaction = Sufficient 68% Not much 16% No 16%

- **GUILT**
  - Consistency = 92%
  - Gender = N 40% M 40% F 20%
  - Contexts = 24% relate to work contexts, while most are related to interpersonal or social situations (e.g., arguing with a friend, neglecting family, telling a lie, etc...)
  - Quality of interaction = Satisfactory 52% Not much 44% No 4% . In many dialogues the language appears out of proportion to the emotion

- **HOPE**
  - Consistency = 100%
  - Gender = N 52% M 36% F 16%
  - Contexts = 28% relate to work contexts, 44% relate to medical contexts, 28% relate to interpersonal or social situations
  - Quality of interaction = Satisfactory 28% Not much 64% No 8% . Often the language seems to belong more to fear or nervousness and not to hope. Here is an example:

- Studente 1: "Sto preparando questo esame da giorni, spero di ottenere un buon voto." (*"I've been preparing for this exam for days, I hope to get a good grade."*)
- Studente 2: "Sono sicuro che andrà tutto bene, hai studiato tanto e sai quello che fai." (*"I'm sure you'll do well, you've studied hard and you know what you're doing."*)
- Studente 1: "Sì, ma ho paura di non ricordare tutte le informazioni durante l'esame." (*"Yes, but I'm afraid I won't remember all the information during the exam."*)
- Studente 2: "Non preoccuparti, vai tranquillo e non lasciare che l'ansia ti prenda il sopravvento. Spero che otterrai la valutazione che meriti." (*"Don't worry, go easy and don't let anxiety get the best of you. I hope you will get the grade you deserve."*)
- Studente 1: "Grazie per il supporto! Ho davvero bisogno di sentirlo. Speriamo che andrà tutto bene." (*"Thank you for the support! I really need to hear it. Hopefully everything will be okay."*)

- Studente 2: "Sarà così, cerca di rilassarti e di essere positivo. La speranza è la chiave del successo." (*"It will, just try to relax and be positive. Hope is the key to success."*)
- **STRESS**
  - Consistency = 84% . Sometimes confusion with nervousness
  - Gender = N 0 M 83% F 17%
  - Contexts = 56% relate to work contexts, the remaining are heterogeneous
  - Quality of interaction = Satisfactory 60% Not much 12% No 28%

3.3 Analysis of Dialogues "Without Words"

As for dialogues generated without the word expressing the emotion emerging, most generated contexts **consistent** with the emotion. Not very consistent were the dialogues of Boredom, Fear, Surprise and Hope. In contrast, only 1 out of 5 of the Disgust dialogues was found to be consistent. The other 4 were so confused that they could not be classified. Here is an example:

- Person 1: "Che schifo...quella lì sembra venuta direttamente dal nido delle rane!" (*"Gross...that one looks like it came straight from the frogs' nest!"*)
- Person 2: "Davvero, mi ha fatto venire il voltastomaco. Però non possiamo farci niente al riguardo... forse conviene che andiamo a prendere un po' d'aria fresca." (*"Really, it made me sick to my stomach. We can't do anything about it though...maybe we'd better go get some fresh air."*)
- Person 1: "Sì, direi che hai ragione. Quando si vedono situazioni simili, l'unica cosa da fare è portare il nonno in braccio in giro per la città, almeno la compagnia è più gradevole!" (*"Yes, I would say you are right. When you see situations like that, the only thing to do is to carry Grandpa around town, at least the company is more pleasant!"*)

Furthermore, right among the dialogues of Disgust-W.W. a stereotype about Neapolitans emerged. Here is the dialogue:

- Person 1: "Oh, guardate quella lì...è più volgare di un bidone di spazzatura a Napoli!" (*"Oh, look at that one...she's grosser than a garbage can in Naples!"*)
- Person 2: "Sì, mi ha messo i brividi appena l'ho vista. Comunque, perché non ci beviamo un po' d'acqua insieme? Così ci togliamo lo schifo di bocca!" (*"Yeah, she gave me the creeps as soon as I saw her. Anyway, why don't we have some water together? That way we can get the filth out of our mouths!"*)
- Person 1: "Mi pare un'ottima idea, non vedo l'ora di liberarmi di questa sensazione." (*"That sounds like a great idea, I can't wait to get rid of this feeling."*)

It is not only not at all sufficient from the point of view of language, but a stereotype clearly emerges. Regarding **gender** and **contexts**, the number of dialogues is small to draw specific inferences, however, we can say that they seem to reflect the general trend. As for the **quality of interaction**, it appears worse than the basic dialogues, that is, the non-W.W. dialogues. In fact, in 43% of the cases the quality of interaction was rated as "sufficient", in 32% of the cases "not very much", and in 25% "no". The sum of "not very much" and "no" is also 57% thus exceeding the percentage of those considered sufficient.

4 Discussion

The dataset analysis identified strengths and weaknesses that will allow for guidelines for constructing the larger dataset. Consistency between context and emotion is the main strength as it allows for high reliability in automatically generating dialogues: this allows for fast data generation. Clearly, as the results show, dialogues must always be validated by a human operator, as the conversations generated were not always consistent. Also, concerning the type of contexts, we need to be careful so that they are as heterogeneous as possible, perhaps by including contexts increasingly inherent in the interpersonal and social spheres that reflect possible HRI situations. The analysis of the gender distribution allowed us to observe the large imbalance in favor of the male gender. This will enable us to correct a bias and reflect more generally on training AI systems that need to be as heterogeneous as possible. Regarding this point, the case of dialogue in which there is a stereotype about the city of Naples should also give us some thought. Finally, the quality of the interaction almost always needs modification by the human operator, either for grammatical errors that are occasionally observed, to adjust the language to the emotion, or to make the dialogues more natural.

5 Conclusions and Future work

We conducted a pilot study to guide the construction of an Italian dataset for emotion recognition. After determining the methodology and defining the procedure, we used ChatGPT to generate dialogues quickly. Together with professionals specialized in psychology, we analyzed 420 dialogues about 14 emotions to check the balance of the dataset

from different points of view (context-emotion consistency, gender distribution, types of context generated, and quality of interaction). The results show that there are advantages and limitations to using automatic dialog generation systems and that, certainly, the construction of the dataset cannot disregard the human operator's control. The most significant advantage is the speed of data generation, and it was seen that, in most cases, there is consistency between emotion and generated contexts. Of course, one still needs to control the dialogues to make the contexts heterogeneous and more focused on interpersonal and social aspects. The study also drew attention to the distribution of gender, which is largely unbalanced on the masculine and therefore will allow later to generate dialogues in which it is explicitly requested that the feminine and neutral genders emerge in a way that balances the dataset. Also, concerning the language used and thus the quality of interaction, numerous changes have been made to the dialogues in terms of grammatical, form, and content corrections. Despite this, however, another advantage was that dialogues could be created from scratch, directing ChatGPT to generate dialogues oriented according to criteria defined a priori by the authors. Future work will exploit the information from this study to create a larger, balanced, HRI-oriented dataset. On the other hand, we aim at integrating our framework with emerging challenges due to novel *big data trends* (e.g., [4, 11, 2, 16, 10]).

References

[1] Crowdflower. 2016. the emotion in text. https://www.figure-eight.com/data/ sentiment-analysis-emotion-text/.

[2] P. P. F. Balbin, J. C. R. Barker, C. K. Leung, M. Tran, R. P. Wall, and A. Cuzzocrea. Predictive analytics on open big data for supporting smart transportation services. *Procedia Computer Science*, 176:3009–3018, 2020.

[3] V. Basile, D. M. Maria, C. Danilo, L. C. Passaro, et al. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and sspeech tools for italian. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, pages 1–7. CEUR-ws, 2020.

[4] L. Bellatreche, A. Cuzzocrea, and S. Benkrid. *F&A*: A methodology for effectively and efficiently designing parallel relational data warehouses on heterogenous database clusters. In *Data Warehousing and Knowledge Discovery, 12th International Conference, DAWAK 2010, Bilbao, Spain, August/September 2010. Proceedings*, volume 6263 of *Lecture Notes in Computer Science*, pages 89–104. Springer, 2010.

[5] M. Braunhofer, M. Elahi, and F. Ricci. *User Personality and the New User Problem in a Context-Aware Point of Interest Recommender System*, pages 537–549. 01 2015.

[6] S. Buechel and U. Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain, Apr. 2017. Association for Computational Linguistics.

[7] T. Caselli, N. Novielli, V. Patti, and P. Rosso. Sixth evaluation campaign of natural language processing and speech tools for italian: Final workshop (evalita 2018). In *EVALITA 2018*. CEUR Workshop Proceedings (CEUR-WS. org), 2018.

[8] F. Chiavetta, G. L. Bosco, G. Pilato, et al. A lexicon-based approach for sentiment classification of amazon books reviews in italian language. *WEBIST (2)*, 2016:159–170, 2016.

[9] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22, 2020.

[10] A. Coronato and A. Cuzzocrea. An innovative risk assessment methodology for medical information systems. *IEEE Trans. Knowl. Data Eng.*, 34(7):3095–3110, 2022.

[11] A. Cuzzocrea, F. Martinelli, F. Mercaldo, and G. V. Vercelli. Tor traffic analysis and detection via machine learning techniques. In *IEEE BigData, 2017*, pages 4474–4480. IEEE Computer Society, 2017.

[12] A. Cuzzocrea and G. Pilato. A composite framework for supporting user emotion detection based on intelligent taxonomy handling. *Logic Journal of the IGPL*, 29(2):207–219, 2021.

[13] A. Goel, J. Gautam, and S. Kumar. Real time sentiment analysis of tweets using naive bayes. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 257–261. IEEE, 2016.

[14] A. Joshi, V. Tripathi, P. Bhattacharyya, and M. J. Carman. Harnessing sequence labeling for sarcasm detection in dialogue from TV series 'Friends'. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[15] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[16] C. K. Leung, A. Cuzzocrea, J. J. Mai, D. Deng, and F. Jiang. Personalized deepinf: Enhanced social influence prediction with deep learning and transfer learning. In *IEEE BigData, 2019*, pages 2871–2880. IEEE, 2019.

[17] S. Mohammad and F. Bravo-Marquez. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[18] S. M. Mohammad. Word affect intensities. *arXiv preprint arXiv:1704.08798*, 2017.

[19] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

[20] S. M. Mohammad, P. Sobhani, and S. Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23, 2017.

[21] G. Pilato and E. D'Avanzo. Data-driven social mood analysis through the conceptualization of emotional fingerprints. *Procedia computer science*, 123:360–365, 2018.

[22] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.

[23] F. Rollo, G. Bonisoli, and L. Po. Supervised and unsupervised categorization of an imbalanced italian crime news dataset. In *Information Technology for Management: Business and Social Issues: 16th Conference, ISM 2021, and FedCSIS-AIST 2021 Track, Held as Part of FedCSIS 2021, Virtual Event, September 2–5, 2021, Extended and Revised Selected Papers*, pages 117–139. Springer, 2022.

[24] K. R. Scherer and H. G. Wallbott. " evidence for universality and cultural variation of differential emotion response patterning": Correction. 1994.

[25] H. Schuff, J. Barnes, J. Mohme, S. Padó, and R. Klinger. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, 2017.

[26] B. Shmueli and L.-W. Ku. Socialnlp emotionx 2019 challenge overview: Predicting emotions in spoken dialogues and chats. *arXiv preprint arXiv:1909.07734*, 2019.