A Multimodal Dataset of Images and Text to Study Abusive Language

Stefano Menini
Fondazione Bruno Kessler
Trento, Italy
menini@fbk.eu

Alessio Palmero Aprosio
Fondazione Bruno Kessler
Trento, Italy
aprosio@fbk.eu

Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
satonelli@fbk.eu

Abstract

English. In this paper, we present a novel dataset composed of images and comments in Italian, created with teenagers in classes using a simulated scenario to raise awareness on cyberbullying phenomena. Potentially offensive comments have been collected for more than 1,000 images and manually assigned to a semantic category. Our analysis shows that the presence of human subjects, as well as the gender of the people present in the pictures trigger different types of comment, and provides novel insight into the connection between images posted on social media and offensive messages. We also compare our corpus with a similar one obtained with WhatsApp, showing that comments to images show different characteristics compared to text-only interactions.[1]

1 Introduction

In order to study abusive language online, the availability of datasets containing the linguistic phenomena of interest are of crucial importance. However, when it comes to specific target groups, for example teenagers, collecting such data may be problematic due to issues with consent and privacy restrictions. Furthermore, while text-only datasets for abusive language detection have been widely developed and used by the NLP community, limitations set by image-based social media platforms like Instagram make it difficult for researchers to experiment with multimodal data. We therefore present a novel corpus containing images and potentially offensive Italian comments and we analyse it from different perspectives, to investigate whether the subject of the images plays a role in triggering a comment.

The data collection was carried out in several school classes, being part of a 'living lab' to raise awareness on cyberbullying and, more generally, on the use of social media by teenagers. The dataset is freely available on Github[2] and, since the comments were collected with the written consent of parents and teachers, they can be freely used for research purposes, without the ethical implications that would derive from using real data posted by teenage users. The images, instead, are released as a ResNet-18 neural network trained on ImageNet, similar to recent NLP works (Kruk et al., 2019), since they were taken from Instagram and cannot be shared as pictures.

2 Related Work

Several datasets have been created to study hate speech, abusive language and cyberbullying. Most of them include single textual comments or threads annotated as being hateful/offensive/abusive or not. For example Reynolds et al. (2011) propose a dataset of questions and answers from Formspring.me, a website with a high amount of cyberbullying content. It consists of 12,851 posts annotated for the presence of cyberbullying and severity. Another resource developed by Bayzick et al. (2011) consists of conversation transcripts (thread-style) extracted from MySpace.com, which are annotated for presence and typology of cyberbullying. For an overview on existing annotation schemes and datasets specific to cyberbullying see the survey presented in (Emmery et al., 2019). Similarly, a project called Hate Speech Datasets[3] (Vidgen and Derczynski, 2020) collects a comprehensive list

[2] https://github.com/dhfbk/creep-image-dataset

[3] https://github.com/leondz/hatespeechdata

of datasets that are annotated with offensive language, online abuse, and so on.

Probably the most popular datasets shared within the NLP community have been extracted from Twitter because of its relatively easy-to-use APIs. Indeed, most of the shared tasks recently organised to build and evaluate hate speech detection systems use Twitter data (Basile et al., 2019; Struß et al., 2019; Bosco et al., 2018; Aragón et al., 2019).

The relationship between textual content and images and the role that they play together is a relatively understudied problem in relation to online hate speech. A notable exception is the dataset collected from Instagram by Hosseinmardi et al. (2015), which consists of 2,218 media sessions, each being annotated with information on cyber-agressive behavior. In this dataset the annotation refers to a thread and not to single offensive messages. The corpus has been used also for classification tasks, for example in (Cheng et al., 2019) it was employed to detect cyberbullying through a hierarchical attention network that takes into account the hierarchical structure of social media sessions and the temporal dynamics of cyberbullying. Other Instagram datasets have been created but they cannot be shared due to the restrictions in the social network policy (Yang et al., 2019).

3 Annotation Tool and Process

The annotation was performed involving overall 95 students aged between 15 and 18. The activity was carried out in classes, during 'living labs' aimed at raising teenagers' awareness on online harrassment and cyberbullying. Students were given access to the CREENDER tool[4], a web-based annotation system that displays pictures taken from a pre-defined batch of images, and allows users to add comments (Palmero Aprosio et al., 2021). In this case, the images were first extracted from Instagram by the authors of this paper and then manually checked to avoid nudity and explicit sexual content.

After a user logs in the system, a picture is displayed, and a prompt asks "*If you saw this picture on Instagram, would you make fun of the user who posted it?*". If the user selects "*No*", then the system picks another image randomly and the same question is asked. If the user clicks on "*Yes*", a second screen opens where the user is asked to specify the reason why the image would trigger such reaction by selecting one of the following categories: "*Body*", "*Clothing*", "*Pose*", "*Facial expression*", "*Location*", "*Activity*" and "*Other*". The user should also write the textual comment s/he would post below the picture. After that, the next picture is displayed, and so on.

The question posed by the system does not ask explicitly whether the user would *insult*, *harrass* or *offend* the person who posted the image, because in a preliminary test with students we observed that the answer would almost always be "*No*". This showed that only in few cases a user would consciously harm another user, especially if the two know each other. Furthermore, comments with explicit hateful content are easy to find online and can be unambiguously annotated in most of the cases. We therefore decided to focus on a more nuanced form of offensive message, that is when a user makes fun of another one. We made this choice because we assumed that this kind of messages would be more ambiguous, containing ironic or sarcastic comments, and mixing humorous and abusive content without being necessarily explicit. This would make the collected data very interesting from a linguistic and computational point of view.

The data collection was embedded in a larger process that required two to three meetings with each class, one per week, involving every time two social scientists, two computational linguists and at least two teachers. During these meetings several activities were carried out with students, including simulating a WhatsApp conversation around a given plot as described in (Sprugnoli et al., 2018), commenting on existing social media posts, and annotating images as described in this paper. Since ethical issues were a main concern since the drafting of the study design, because all participants were underage students, all the activities had been co-designed with the schools involved and informed consent was gathered beforehand both from teachers and from parents.

The sessions with students were organised so that different school classes annotated the same set of images, in order to collect multiple annotations on the same pictures. However, since some users were quicker than others in giving a judgement on the pictures, we could not collect multiple annotations for all images included in the dataset (see

[4] https://github.com/dhfbk/creender/tree/master

Pictures with ↓ ... and having →	At least 1 comment	(Total comments)	No comments
At least 1 judgement	1,018	1,135	16,894
At least 2 judgements	901	1,018	9,876
At least 3 judgements	713	815	5,454
At least 4 judgements	495	563	3,060

Table 1: Number of pictures in the dataset with at least *n* judgments (*'yes'/'no'*) and number of comments.

Table 1 for details).

4 Annotated Corpus

Overall, 17,912 images have been judged at least once by the students. For 1,018 of them, at least an offensive comment has been written during the annotation sessions. Overall, the number of comments in the dataset is 1,135, which is higher than the number of pictures with a comment because the same image may be commented more than once by different students. An overview of the content of the dataset including images and comments is presented in Table 1. Note that the number of *judgements* refers to the 'yes/no' option selected by users in the first platform view, while the number of *comments* refers only to the images tagged with a 'yes', for which a student wrote also a comment. Overall, only one image has been tagged with four 'yes', and in most of the cases annotators selected only 'no'. The number of images tagged with exactly three 'yes' is 13, those with two is 88. Since these images have been leveraged from Instagram with no particular criterion in mind, the distribution of 'yes' and 'no' may be considered realistic, with the majority of pictures not triggering any potentially negative reaction, and around 6% of them being associated with offensive comments.

In general, we observe that there is a low agreement on whether a picture triggers an offensive comment or not. This suggests that an offensive intent is more dependent on the attitude of a user posting a comment than on image-specific features. We also compute inter-annotator agreement – using Krippendorff's alpha measure (Krippendorff, 1970) – on the trigger categories assigned to the comments, considering only the images that received at least two comments. Agreement is 0.19, which implies again that the reason to make fun of a user does not depend on a specific feature of the picture, but rather that multiple aspects of a posted image can be taken as an excuse for

	Female	**Male**	**Both**	**None**	**Total**
Body	27	20	3	4	54
Clothing	66	30	9	12	127
Pose	114	99	11	5	229
Facial Exp.	68	90	17	7	182
Location	16	17	7	57	97
Activity	12	14	7	36	69
Other	72	63	22	113	272
Total	377	318	76	252	1023

Table 2: Distribution of offence triggers per subject types

potentially offensive comments. In order to avoid the ambiguity introduced by the *'Other'* label, we also compute IAA ignoring this class. This time the agreement value is 0.26, showing that on the one hand the 'Other' label covers uncertain cases, but also that the reason to comment a picture remains highly subjective.

Some typical comments collected during the simulation are *Copriti (Cover yourself up)*, *Che schifo di foto (This picture sucks)* and *Inquietante (Disturbing)*. These comments have different features compared to hate speech messages extracted from Twitter: they tend to be short because they complement the image and they are rich in deictic expressions. In most of the cases, they are not self-contained from a semantic point of view.

5 Corpus Analysis

In order to analyse whether what is portrayed in a picture has an impact on the choice to write an offensive message, we manually assign each image with at least 1 comment to one of the following categories: male-only subject(s), female-only subject(s), mixed group, no human subject. How the different categories are distributed, taking into account also the triggers (i.e. self-declared reasons to write a comment) is displayed in Table 2.

The last column of the table shows that the categories are rather imbalanced, with a minority of comments associated with the 'Body' label and several comments concerning the 'Pose'. How-

ever, for most of the comments the 'Other' label was used. When collecting feedback from students after the annotation sessions, several annotators suggested that it should be possible to assign multiple labels instead of just one, and reported that they used the 'Other' label for those cases. On the other hand, they did not express the need to include additional categories in the annotation.

As regards the picture subjects, the analysis shows that the main differences between pictures with male and female subjects concern the 'Facial expression' and 'Clothing' categories: the first is more frequently associated with male subjects, while the second seems to be more related to female subjects. When there are multiple subjects with different genders, instead, no particular differences are observed. As expected, when no person is portrayed in the picture, 'Location', 'Activity' and 'Other' are prevalent. In some cases, 'Pose' or 'Expression' are selected, because of the presence of animals or drawings in the images.

We manually assign the subject category also to a set of 3,200 pictures randomly taken from the images that were tagged with 'No'. Then we compare the two category distributions, that are reported in Table 3. By applying the χ^2 test ($N = 4,218$), we observe a statistically significant difference between the two distributions of categorial variables ($p < .001$). In particular, pictures with no human subject are less likely to get an offensive comment, while those with a female subject are the most commented ones. Also male subjects, however, trigger offensive comments very frequently, while they are only present in 19% of the images which were not commented by users.

	% Yes	% No
Females	36.85	32.14
Males	31.09	19.00
Mixed	7.43	9.33
Nobody	24.63	39.53

Table 3: Subject types for pictures annotated with 'Yes' (i.e. triggering a comment) and 'No'

6 Dataset comparison

In order to better understand the peculiarities of the textual comments in our corpus, we compare them with the messages in another existing corpus created with a similar approach, i.e. simulated scenarios, and with the same goal, i.e. study how teenagers communicate online. More specifically, the second corpus was created following the approach described in (Sprugnoli et al., 2018) using WhatsApp chats in classes to simulate cyberbullying interactions among teenagers. The target age group is the same as for our multimodal corpus, but in the second corpus the interactions are solely based on text. We select from the WhatsApp corpus the 3,004 comments manually tagged as offensive so to make them comparable with the 1,135 comments in our multimodal corpus, which were all written with the goal to make fun of someone. Both corpora are processed with the TINT suite (Aprosio and Moretti, 2018), through which a number of linguistic features were extracted.

As regards type/token ratio, it is 0.62 in the WhatsApp corpus and 0.82 in our data, suggesting that images may foster a richer, more creative use of the language, even if offensive. This difference may also be affected by the fact that WhatsApp chats followed a pre-defined plot, therefore limiting the topics to be mentioned in the interactions. Also lexical density is different, being 0.56 on WhatsApp and 0.65 in our corpus. This confirms that the language used in image comments is more complex and more similar to written standard language, while WhatsApp chats share some features of spoken interactions, where content is generally sparser (Stubbs, 1986). We also analyse the impact of nominal utterances over the corpus, counting how many turns do not contain any verb based on the PoS tagger output. While in the WhatApp corpus these utterances are around 29%, in our multimodal corpus they are 35%. According to previous studies (Comandini et al., 2018) this kind of construction is used to express emphasis, and is particularly frequent on social networks and in spoken language. Our results are in line with previous findings on social media language, but show also that in our multimodal corpus the presence of images may boost communicative economy, making verbs less necessary than in other text-based media.

Concerning message length, both corpora contain rather short messages, with 5.9 tokens per sentence in the WhatsApp data and 5.3 tokens in the multimodal corpus on average. The standard deviation is rather high in both cases (4.80 and 4.99 respectively) because of the high length variability of the messages, ranging from one word (e.g. *Copriti*, *Certo*) to max. 50 tokens per message in our

multimodal corpus and 67 in the WhatsApp one.

Question marks are abundant in the WhatsApp dataset (0.51 per sentence on average, vs. 0.19 in the other corpus) because it contains interactions including questions and answers. Exclamation marks, instead, are much more frequent in the multimodal corpus (0.14 per sentence vs. 0.03), which contains more emphatic comments.

7 Conclusions

In this work we present a multimodal dataset in the abusive language domain created by teenage participants who, during online annotation sessions, judged whether images may trigger an offensive comment, left a possible comment and also assigned to it a trigger category.

The analysis of the collected data gives interesting insights into how Instagram-like platforms work. First of all, images containing persons are more likely to trigger potentially offensive comments than those without a human subject. Both female and male subjects are offended but the reasons may differ: the former are targeted more because of the pose and of the clothing, while the latter for the pose and the facial expression. In general, the reasons why a comment is triggered seems to be subjective, depending on the user leaving the comment rather on some actual characteristics of the person portrayed in the picture.

We conducted our data collection using Instagram pictures randomly taken from this platform, because it is the social network that is most used by teenagers, including those involved in our annotation sessions. However, this makes the release of the full dataset impossible, because of a very restrictive policy concerning images. We therefore adopt a strategy already used within the NLP research community (Kruk et al., 2019), releasing the images as a layer of a ResNet-18 neural network trained on ImageNet. The comments, instead, are freely available without restrictions due to the consent signed by all parents and by the anonymity granted to participants. This represents a very interesting dataset from a research point of view, since it includes comments written by underage students that are usually difficult to obtain because of privacy reasons.

In the future, we plan to extend our study to compare the judgements given by single users to those given by groups of peers. In a preliminary study, we observed that, when students are given the possibility to discuss with a small group of peers whether they would like to write an offensive comment, they tend to be more aggressive and are more likely to select 'yes'. While the comments collected so far with groups of annotators are not enough to allow a fair comparison between the two settings (single vs. group), we plan to extend them in the future and pursue also this interesting research line. Finally, we plan to train a classifier able to detect offensive messages by merging visual and textual features with the goal to integrate it in a monitoring tool like the one introduced in (Menini et al., 2019). This would enable a more holistic, context-aware understanding of offensive communication online.

Acknowledgments

Part of this work has been funded by the KID_ACTIONS REC-AG project (n. 101005518) on "Kick-off preventIng and responDing to children and AdolesCenT cyberbullyIng through innovative mOnitoring and educatioNal technologieS". We would like to thank the students involved in the corpus creation process and their teachers who supported the experimentation.

References

Alessio Palmero Aprosio and Giovanni Moretti. 2018. Tint 2.0: an All-inclusive Suite for NLP in Italian. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mario Ezra Aragón, Miguel Álvarez-Carmona, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Daniela Moctezuma. 2019. Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SEMEVAL 2019*, Minneapolis, Minnesota, USA, June.

Jennifer Bayzick, April Kontostathis, and Lynne Edwards. 2011. Detecting the presence of cyberbullying using computer software. In *3rd Annual ACM Web Science Conference (WebSci 2011)*, pages 1–2.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of EVALITA 2018*.

Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 235–243. SIAM.

Gloria Comandini, Manuela Speranza, and Bernardo Magnini. 2018. Effective Communication without Verbs? Sure! Identification of Nominal Utterances in Italian Social Media Texts. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Chris Emmery, Ben Verhoeven, Guy De Pauw, Gilles Jacobs, Cynthia Van Hee, Els Lefever, Bart Desmet, Véronique Hoste, and Walter Daelemans. 2019. Current Limitations in Cyberbullying Detection: on Evaluation Criteria, Reproducibility, and Data Scarcity.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishr. 2015. Prediction of cyberbullying incidents on the Instagram social network. *arXiv preprint arXiv:1508.06257*.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4621–4631.

Stefano Menini, Giovanni Moretti, Michele Corazza, Elena Cabrio, Sara Tonelli, and Serena Villata. 2019. A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 105–110.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2021. The CREENDER Tool for Creating Multimodal Datasets of Images and Comments. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59. Association for Computational Linguistics.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

Michael Stubbs. 1986. Lexical density: A technique and some findings. In Michael Coulthard, editor, *Talking about Text. Discourse Analysis*, chapter Lexical Density: A Technique and Some Findings, page 27–42. English Language Research.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670*.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 11–18, Florence, Italy, August. Association for Computational Linguistics.