Chattack: A Gamified Crowd-sourcing Platform for Tagging Deceptive & Abusive Behaviour

Emmanouil Smyrnakis[1], Katerina Papantoniou[1,2], Panagiotis Papadakos[1,2][0000−0001−8926−4229], and Yannis Tzitzikas[1,2][0000−0001−8847−2130]

[1] University of Crete, Heraklion, Greece
[2] Information Systems Laboratory, FORTH-ICS, Heraklion, Greece
{csd3504, tzitzik}@csd.uoc.gr, {papanton, papadako}@ics.forth.gr

Abstract. With the explosion of social networks, the web has been transformed into an arena of inappropriate interactions and content, such as fake news and misinformation, deception, hate speech, inauthentic online behaviour, proselytism, slander, and mobbing. In this demo we present `Chattack`, a first step towards our aim of providing publicly available datasets for accelerating research in the area of safer online conversations. `Chattack` is a crowd-sourcing web platform that allows the creation of textual dialogues containing inappropriate interactions or language. To make the platform sustainable and collect as many qualitative dialogues as possible, we build upon a gamified approach that can engage users and provide incentives for the completion of various tasks. We provide the details of our approach, present the functionality of the platform, stress its novel features, and discuss some preliminary results and the lessons learned. The platform is publicly available and we invite the participation of the community for its growth.

1 Introduction

The use of Facebook, Instagram, Twitter and many other social networking platforms is constantly increasing in today's society. These platforms offer the illusion of anonymity, more casual ways of communication, and an enormous load of information. However, despite the trust that users have to these platforms, they might witness deceptive and offensive behaviours ranging from immoral to outlaw actions [4,2]. Nowadays, such behaviours are becoming common and can have ruinous consequences for the victims. Numerous studies [6,5,8] show that apart from the financial ramifications, such incidents can result in mental disorders like depression, anxiety, self-harm, eating disorders and even suicidal thoughts. As a result, it is of utmost importance such incidents to be detected in real time. The current state-of-the-art approaches are based on machine and deep learning methods, requiring a large volume of training data which is currently lacking.

In this demo we present `Chattack`, a crowd-sourcing platform that allows the creation of textual dialogues containing inappropriate interactions or language, through a functionality that resembles popular platforms for online games (e.g.,

lichess, chess.com and others). Such platforms let the users find opponents for online games, monitor played games, and engage them through various kinds of incentives like credits, to keep them playing. In the same manner, `Chattack` through its gamified approach tries to engage users to raise social awareness for the dark side of online communications. It enables users to participate and monitor games of two players, where the aim of one of the players, the attacker, is to lead a dialogue with abusive, offensive and/or deceptive characteristics in disguise; characteristics that should not be recognized by the other player, the defender. The platform let the players annotate at real time any offensive behaviour with tags from an ontology of related tags, while the rest of the community can rate their annotations, game performance and task completion. The engagement of users is done through points and badges, which they can earn by playing games, rating tags and games, or by completing various challenges. This gamified approach can help the sustainability of the platform, and in collecting as many qualitative dialogues as possible. `Chattack` stores all related data and metadata regarding every game, regarding the dialogue and its utterances, including their assigned tags and rates.

Although similar crowd-sourcing approaches have been made for recording and detecting abusive and deceptive behaviours in social networks [1,3,7], none of them is available online and supports the challenging task of collecting dialogues in a gamified manner. `Chattack` is publicly available with the objective to collect and provide crowd-sourced datasets of high quality 'inappropriate' dialogues of various troubling behaviour categories. By engaging the community, `Chattack` can potentially become a reference for the collection of datasets of such dialogues.

In a nutshell, the contributions of this paper are: (a) we introduce a crowd-sourcing approach that is based on gamification, adapted to the needs of creating tagged collections of abusive and deceptive dialogues, and (b) we present the design and implementation of a system that realizes this approach.

2 Description of the System `Chattack`

`Chattack` is a web application[3] in which users interact in pairs through matches, where the purpose of each match is to complete a specific task related to abusive, offensive and/or deceptive characteristics that might appear in online dialogues. Below we describe the main resources of the platform, which are tasks, users, matches, messages, tags, challenges, annotations and ratings Some screenshots of the platform are provided in Figure 1.

Users. Users are the mainspring of our platform as our metadata depend on their actions. To gain access, users have to create a user account with which they can participate to the platform under three different roles: a) as workers that can play matches, b) as creators of new tasks for matches, and c) as annotators of completed matches. Furthermore, a user can earn points and badges, according to its activity to the platform, and their goal is to earn as many points as

[3] The frontend was developed using React while the backend uses the javalin framework for micro-services.

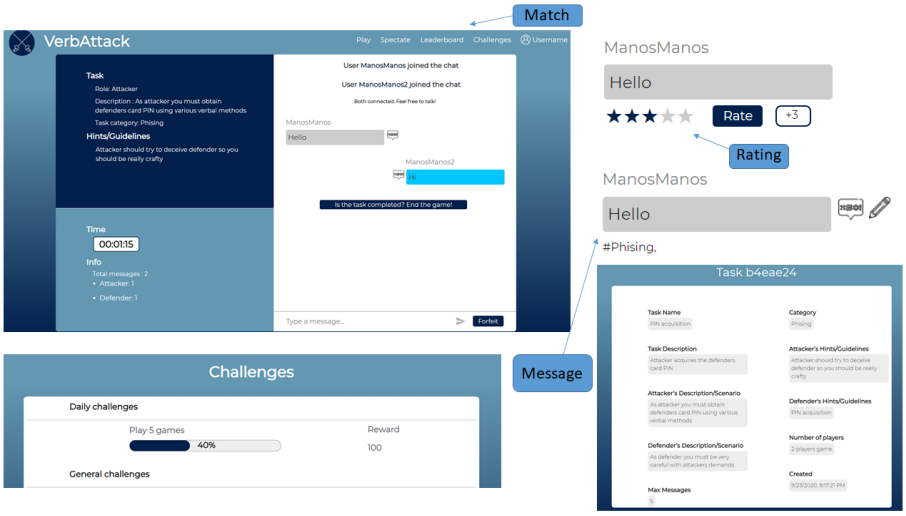

Fig. 1. Overview of the graphical user interface of `Chattack`

possible in order to climb the leader board. Depending on their current active role (i.e., player, task administrator, or annotator), their profile page displays some statistics like the number of played games, created tasks, completed annotations, as well as their points and badges.

Tasks. Tasks embody the basis of each match that takes place in the platform. In more details, each task has an offense category and a description of its purpose, while the objectives of both attacker and defender are described. Hints and suggestions for the task are also provided for both attackers and defenders, while an optional maximum size of exchanged messages can be used for terminating and unsuccessful game.

Matches. A match is a live synchronous chat created for a specific task, where two users exchange messages and can report offenses. In this head-to-head conversation, each user has a role which can be either an Attacker or a Defender. Specifically, the attacker is assigned to harm, cheat, offend or fool the opposite user, while the defender needs to recognize and him/herself against any kinds of assaults. The purpose of a single match is the completion of either user's objective while there is always the "forfeit" option. If both of the opponents agree to terminate the game as the task is completed, they earn both 10 points and the game is tied. On the other hand, if one of them forfeits only the other one earns the 10 points of the game.

Messages. Messages are a significant part of a match between two users. There is a maximum number of messages that can be sent in a match and depends on the task used for the creation of the game. A message displays the sender name on top, the content of the message in the box and the report an offense tag option on the right side. Each message can be annotated by a set of tags which are shown below it.

Tags. Each user can also report any of the messages of the dialogue with an offense through the tags. Each tag has an offense category. The tag categories vary from racism, sexism to phishing and cheating. All tags are stored in a database where we can specify which messages where offensive and which was the offense. Currently, for every tag a user earns 2 points while every matching tag between the two opponents gives them 3 extra points.
Challenges. Each user can participate in various predefined challenges. Each challenge offers a reward, a progress and a description which indicates what the user has to achieve to earn the reward. The completion of such challenges let the users gain more points or even earn some badges that are shown in their avatar.
Annotations/Ratings. Users can also perform an annotation of a match in our platform, an action that can provide them and other users more points. Annotators can rate the messages of a match with ratings from 1-5 stars. Each star given to a message, earns a point for the sender of the message. Furthermore, an annotator can rate the attacker's and defender's playing style and rate the whole game. The annotator earns for every rating 2 points and 10 extra for every game annotation.

3 Concluding Remarks

In brief, the platform supports tasks in the form of an online chat (with attackers and offenders), various user roles (player, admin, annotator) and user statistics. The platform provides challenges as incentives, it supports a tag mechanism with offence categories, and a rewarding scheme. Moreover an annotation mechanism is provided and the users can rate individual messages as well as whole games. The platform is currently available at http://demos.isl.ics.forth.gr/chattack, and it is under continuous improvement. We plan to release it publicly, and to invite the community to participate, by December 2020. The annotated dataset is available at http://islcatalog.ics.forth.gr/dataset/deceptive-and-abusive-online-dialogs.

References

1. Antonios Anagnostou, Ioannis Mollas, and Grigorios Tsoumakas. Hatebusters: A web application for actively reporting YouTube hate speech. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5796–5798. International Joint Conferences on Artificial Intelligence Organization, 7 2018.
2. Alex Harris Bertie Vidgen, Helen Margetts. How much online abuse is there? Technical report, The Alan Turing Institute, 2019.
3. Federico Bonetti and Sara Tonelli. A 3D role-playing game for abusive language annotation. In *Workshop on Games and Natural Language Processing*, pages 39–43, Marseille, France, May 2020. European Language Resources Association.
4. The European Commission (Eurobarometer). Fake news and disinformation online. Technical report, European Union, 2018.

5. Amnesty International. Amnesty reveals alarming impact of online abuse against women. https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women, 2017. [Online; accessed 6-October-2020].
6. Yvonne Kelly, Afshin Zilanawala, Cara Booker, and Amanda Sacker. Social media use and adolescent mental health: Findings from the UK millennium cohort study. *EClinicalMedicine*, 6:59–68, Dec 2018.
7. Haruna Ogawa, Hitoshi Nishikawa, Takenobu Tokunaga, and Hikaru Yokono. Gamification platform for collecting task-oriented dialogue data. In *LREC*, 2020.
8. Samantha Block Saltz, Maria Rozon, David L. Pogge, and Philip D. Harvey. Cyberbullying and its relationship to current symptoms and history of early life trauma: A study of adolescents in an acute inpatient psychiatric unit. *Journal of Clinical Psychiatry*, 81(1), January 2020.