An Agent for Deception Detection in Discussion Based Environments

(Extended Abstract)

Amos Azaria
Dept. of Computer Science
Bar Ilan University, Israel

Ariella Richardson
Dept. of Industrial Engineering
Jerusalem College of Technology, Israel

Sarit Kraus
Dept. of Computer Science
Bar Ilan University, Israel

ABSTRACT

Autonomous agents can be of assistance in detecting and reducing deception in computerized forums and chat-rooms. We focus on text-based environments where the deceiver is a member of a group which is holding a discussion. Deception detection methods which currently exist for such environments, heavily rely on either audio or visual information. We have developed DIG, an innovative machine learning-based autonomous agent, which joins a group of players as a regular member and assists them in catching a deceiver. We introduce "the pirate game" as a platform for deploying this agent. Our experimental study shows that although humans display difficulty detecting deception, DIG is not only capable of finding a deceptive player, it also helps increase the entire group's success.

Categories and Subject Descriptors

I.2.m [**Computing Methodologies**]: ARTIFICIAL INTELLIGENCE—*Miscellaneous*

Keywords

Deception Detection, Human Modeling, Discussions

1. INTRODUCTION

Many activities in our everyday lives involve sharing opinions with peers through computer-mediated communication. People participate in forum discussions on important topics such as: how to raise their children, what medication they should use and how to improve their business. It would be nice to assume that all of the people participating in these discussions have a common, honest goal and that malicious participants are spotted by moderators. However, this is often not true. Pedophiles manage to infiltrate kids' chat rooms, commercial products are pushed in forums by dealers posing as regular users, and business forums are probably full of advice that actually assists their competitors. Computer-mediated communication has provided a modern venue for deception [2], where measuring the extent of such deception is also a topic of active research [1].

Appears in: *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), Lomuscio, Scerri, Bazzan, Huhns (eds.), May, 5–9, 2014, Paris, France.*

We investigate scenarios where several participants attempt to collectively detect a deceptive member. We therefore designed a game for a text-based discussion environment. In this game there are several credible participants and one dishonest participant (a pirate). In the first phase of the game the participants conduct a textual discussion with an attempt to uncover the liar, and later they cast their votes as to whom they think the liar is.

We developed the Deception In Group detector and catcher Agent (DIG). The DIG agent is capable of participating in the game while posing as a regular player. During an ongoing session of the game, DIG uses input from multiple previous games, along with input from the current game, and outputs the next sentence it wants to contribute to the discussion. Although the agent has no enforcement capabilities in the forum, it participates in the group decision of who the liar is and is also able to raise the awareness of other participants to malicious and dishonest activity. This is useful as it implies that users without administrative privileges can deploy an agent into an active chat environment. We use machine learning on the data collected from human participants to determine whether a player is honest or not. The agent uses this information to catch the pirate. We also apply machine learning methods in order to learn when players fall under suspicion. This information is also used by our agent in order to minimize the suspicion that it raises. We focus on the discussion dynamics such as tendencies towards accusation, denial or agreement.

In this short note we provide a platform for deception detection within a text based environment that supports group discussion. In this environment we deploy an agent that is required not only to detect deception and lead other participants to recognize the deceiver, but also to refrain from raising suspicion itself. Previous approaches use corpora and do not have multi-player interaction. To the best of our knowledge we are the first to deploy an autonomous agent in any such environment. We provide evaluation of our agent.

2. THE PIRATE GAME

In order to simulate the environment which we are interested in studying (deception in chat-rooms and forums), we need a game which will provide us with the following properties: 1) The game is played by a group of people. 2) The game uses text-based communication. 3) The game is based on a discussion, using short messages and in which players refer to one another. 4) The deceiver has some motivation

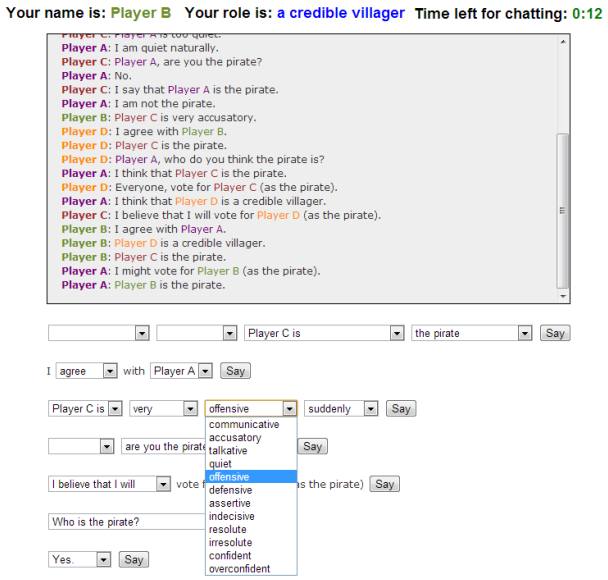

Figure 1: A screen-shot of the pirate game in progress (slightly altered for best fit)

to deceive. 5) The other players have some motivation to find the deceiver.

To satisfy these properties we introduce the Pirate Game; a game, with four players and two roles. Three players are the honest players, the "credible villagers", and the fourth player plays the deceptive participant, the "pirate". All players are informed of their own role but not of anyone else's. The participants are told that they are a group of villagers who went on a journey to find a treasure. They have found a treasure of coins and can split it. However, one of the participants is a pirate and can steal the coins unless he is detected. In order to detect the pirate a discussion phase is held. After the discussion, all credible villagers cast votes as to whom they think is the pirate (or an empty vote if they wish). In our game we do not allow the pirate to vote, as allowing the pirate to vote would make the game harder and less fun for the players and demotivate them. The votes are concealed until all players cast their votes. If there is a majority of votes for the pirate he is "caught" and the money is split between the credible players. Otherwise the pirate receives all the money. At the beginning of the game each player is told his role and the discussion phase begins. The discussion is composed of structured sentences (examples are presented in Figure 1). The interface allows the composition of approximately $4,000$ sentences.

We also implemented a second variation of the game. This variation is different only for the scenario where the pirate manages to escape (does not receive a majority of the votes). In this variation the pirate is told that one of the credible villagers will turn him over to the village ruler, if he escapes with the money. This results in neither the pirate nor the other players receiving any money, unless the pirate manages to convince the other players to cast at least one vote against the villager. In this case the villager will be considered unreliable (to the village ruler) and the pirate will gain all of the coins, that he escaped with. This setting encourages the pirate to be active in the game. We call this version

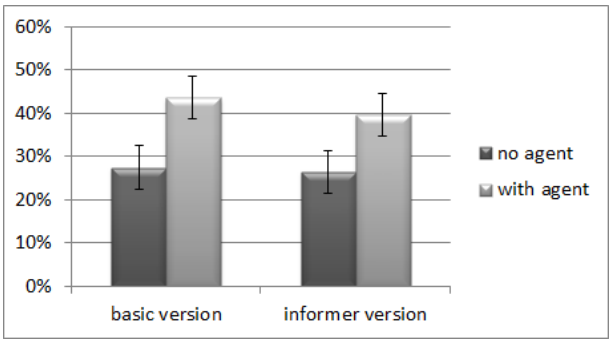

Figure 2: Success rate in catching the pirate. Compares both versions of the game, with and without an agent.

of the game the "informer version" and differentiate it from the "basic version".

3. EXPERIMENTS

Participation in all experiments consisted of a total of 320 subjects from the USA (recruited via Amazon's Mechanical Turk service), of which 47.8% were females and 52.2% were males. The subjects' ages ranged from 18 to 67, with a mean of 32 and median of 30. We ran experiments with the two versions of the game ("informer" and "basic"), each with two different setups. We ran the game with only human players and then with an agent playing the role of one of the credible villagers (the agent is never the pirate). The subjects weren't told about the agent and therefore assumed all players were humans. According to comments we collected, no players suspected a nonhuman player.

Figure 2 presents the success rate of the credible villagers at catching the pirate in both versions of the game, with and without the agent. In both versions of the game the groups including the DIG agent (basic:43.7%, informer:39.7%) significantly outperform (using chi square test, with $\alpha = 0.05$) the groups that didn't include the DIG agent (basic:27.5%, informer:26.5%) . The performance of the human players without the agent is very close to the expected utility of random voting which is 0.26.

4. ACKNOWLEDGMENTS

This work was supported in part by ERC grant #267523, MURI grant number W911NF-08-1-0144 and ARO grants W911NF0910206 and W911NF1110344.

5. REFERENCES

[1] M. Ott, C. Cardie, and J. Hancock. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 201–210, New York, NY, USA, 2012. ACM.

[2] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated. *Group Decision and Negotiation*, 13:81–106, 2004.