What Would Be the Principles for Successful Trollbot Design?

Heidi Vepsäläinen
heidi.vepsalainen@helsinki.fi
Department of Computer Science,
University of Helsinki
Helsinki, Finland

Henna Paakki
Department of Computer Science, Aalto University
Espoo, Finland
henna.paakki@aalto.fi

Antti Salovaara
Department of Design,
Aalto University
Espoo, Finland
antti.salovaara@aalto.fi

ABSTRACT

As far as we know, trollbots that would be indistinguishable from humans and would succeed in luring people into endless frustrating conflicts without being recognized as bots do not yet exist in social media. It is though very likely, that there is a desire to design one for malicious purposes. Here we speculate on the idea of designing a successful trollbot for research purposes by using concepts that derive from Conversation Analysis and Natural Language Framework. Based on our ongoing reseach on trolling, we argue that a successful trollbot would need to prevent its interlocutor from reaching their goal in a given context, but at the same time manage to keep the other party expecting that they would be able to reach a common ground at some point.

CCS CONCEPTS

• **Human-centered computing** → **Social media**.

KEYWORDS

chatbot, Conversation Analysis, Natural Language Framework, trollbot

2021 ACM CHI Conference workshop “Let’s Talk About CUIs: Putting Conversational User Interface Design Into Practice”.

INTRODUCTION

While social media have created a multiplicity of opportunities for social contacts, they have also created a whole new array of challenges. One of the challenges is online trolling: a class of behaviors that cause anxiety and waste of time among people on online platforms, as well as possible exposure to misinformation. Internet trolling appears often as deceptive provocations in several forms, from overt displays of aggression and shocking to harder-to-recognize covert strategies such as digression to irrelevant matters or expressions that increase antipathy and polarization [1]. A further concern is that soon socially competent conversational agents (i.e., "chatbots") may participate in social media disguised as humans. Such agents may act like trolls, and when deployed in masses, may also create false impressions of polarization or consensus among the humans. This offers new means to advance the chatbot creators' agenda. By far trollbots – as far as we know – have not been very successful in either deceiving people into thinking they are humans or luring people into endless discussions.

We will utilize concepts from Conversation Analysis (CA), CA-based Natural Conversation Framework [4], and research on natural trolling to hypothesize, what would a successful trollbot look like. CA-based design methods could potentially lead to more socioculturally contextual and thus engaging designs [2], and, in the case of trollbots, create more frustration and distortion.

DESIGNING A TROLLBOT

Why would a researcher want to design a trollbot?

Our interest in building troll bots rises from our current project Automated trolling and fake news generation in future social media: computational and empirical investigations of the threat and its implications. We acknowledge that there are risks involved in a study that even hypothesizes with building a successful trollbot, as the findings could be used for sinister purposes. There are, though, two important reasons why researchers could and should be interested in the creation of troll bots. 1) To investigate in a controlled setting what kind of harm a trollbot that passes as a human could do, we need a trollbot that passes as a human. 2) As we are likely not the only ones attempting this, and there might already be social bots that are designed for malign purposes and are indistinguishable from humans, research on the topic is of utmost importance in defining the countermeasures against these bots.

Hypotheses and principles for trollbot design

To design a successful trollbot, that is, one that could deceive humans, one needs to have 1) a definition of trolling, 2) understanding or at least a strong hypothesis on how trolling is conducted, and 3) a hypothesis on how this behavior might be mimicked.

Table 1: A discussion pattern between virtual agent (A) and user (U), according to Moore & Arar [4, p. 231].

Pattern		
01	U:	PARTIAL REQUEST
02	A:	DETAIL REQUEST
03	U:	DETAIL
04	A:	DETAIL REQUEST
05	U:	DETAIL
06	A:	DETAIL REQUEST
07	U:	DETAIL
08	A:	HOLD REQUEST
09	A:	GRANT
10		<OFFER OF ARTIFACTS>
11	U:	SEQUENCE CLOSER
12	A:	RECEIPT

Example		
01	U:	I want to book a flight
02	A:	Okay. Where do you want to go?
03	U:	Kona, Hawaii
04	A:	Where are you leaving from?
05	U:	SJC
06	A:	When would you like to go?
07	U:	June 17–30
08	A:	One moment please...
09	A:	Okay. Here are flights from SJC to KOA on June 17–30
10		((visual flight cards))
11	U:	thank you
12	A:	You're welcome!

Following Hardaker [1] we define trolling as "deliberate (perceived) use of impoliteness/aggression, deception and/or manipulation in CMC to create a context conducive to triggering or antagonizing conflict, typically for amusement's sake." The troll lures others into circular discussions, wasting their time and effort and causing distress.

To hypothesize what are the preconditions for trolling in chat, we can make an analogy into another type of trolling done via telephone. If one were to troll a telemarketer and waste their time without actually buying anything, one needs to consider two things. First, the telemarketer needs to be disoriented with irrelevant talk which allows time to pass so that the telemarketer never reaches their goal of finishing their marketing talk. Secondly, this has to be done discreetly. If the sidetracking talk is too irrelevant, the telemarketer will notice that they are not going to be able to sell anything and close the call. Thus, the telemarketer has to be kept under the impression that they will eventually reach their goal.

The example of wasting a telemarketer's time by preventing them from reaching their goal helps us to understand what one needs to do to troll successfully. We conceptualize it by building on the conversation analytic (CA) concept of conversational projects: courses of actions that the participants are pursuing [3]. Using this concept, participants' goals can be studied using CA's analytical tools. Interestingly, trolling is a context for two competing projects: here one participant – the troll – is not aiming at mutual understanding while the other participants aim to pursue a discussion about a topic important to them.

Conversation Analysis is a method for microanalysis of conversational actions and sequences, and its findings have been successfully implemented in designing conversational agents, like the one in Example 1 [4]. Our proposition for a trollbot design utilizes the CA-based Natural Conversation Framework (NCF) [3], which is built on expandable sequences that carefully mimic human-to-human interaction, especially in customer service situations. As conversational agents, though, aim at reaching common ground with the user and assisting them in achieving their goal, a trollbot built with the same framework would aim at breaching common ground and preventing the user from achieving their goal.

What would a trolling virtual agent look like?

In Example 1 we see a prime example of a virtual agent at work, finding and booking flights for the user. The user and agent work together towards a goal, and after the goal has been reached the exchange is terminated. Thus, the virtual agent serves the user's project [3].

Example 2, in contrast, shows an example of what a trolling virtual agent could look like. As Natural Conversation Framework does not contain intents for trolling, they had to be invented for this example. The virtual agent begins an irrelevant discussion on the user's desired destination, Hawaii, instead of proceeding with booking the tickets. While talking about Hawaii, the trolling virtual agent is topically coherent but does not advance the user's project as expected. We suggest that if the trollbot would say something entirely irrelevant or for example insult the user, the discussion would be terminated faster and thus the troll bot would not be as successful as it is when the user still has some hope for achieving their goal.

In reality, there is probably not a high risk that anyone would be building a trolling virtual agent. The use of a bot like that would be very limited and would not be able to cause a lot of harm. A more likely scenario would be a debating social bot that takes part in forum discussions or chats.

Table 2: A modified hypothetical discussion pattern between a user (U) and a trollbot agent (T).

Pattern		
01	U:	PARTIAL REQUEST
02	T:	DETAIL REQUEST
03	U:	DETAIL
04	T:	EXPLANATION REQUEST
05	U:	EXPLANATION GRANT
06	T:	INFORMATION CHECK
07	U:	DISCONFIRMATION
08	T:	INFORMATION CONFIRMATION
...		

Example		
01	U:	I want to book a flight
02	T:	Okay. Where do you want to go?
03	U:	Kona, Hawaii
04	T:	Why would you want to go there?
05	U:	For a holiday
06	T:	Did you know there are a lot of hurricanes in Hawaii?
07	U:	No..
08	T:	Well you should
...		

What would a debate trollbot look like?

As we start from a presumption that a troll, or at least a subsection of trolls, aims at preventing their interlocutors from reaching their goal, we need to have an understanding of what people want to achieve in different kinds of discussions. For example, forum discussions differ in their purpose, which also affects what kinds of turns people take in them. Some discussions are seeking advice, others are asking for opinions, some might announce an achievement, etc. If we take as an example a political debate that rises from a news article, people generally come there to state an argument of some kind. This argument may be disputed with a counterargument or discussed, and others may offer or request further evidence on the matter. Arguments may be challenged but eventually, everyone is allowed to state their argument even when others disagree with them.

Trolling in debate situations needs to be systematic and last for some time before it can be spotted as trolling. In this given framework for trolling, the trollbot's turns may consist of for example challenges and counterarguments without accounts, which leads the user to think that they may be able to get their argument through by explaining and giving more evidence, as in Example 3. Instead of providing a solid argument and taking part in the debate, the trollbot continues systematically and relentlessly challenging, and the discussion will end only when the user is ready to give up and stop responding. In a best-case scenario, this may lead to a long discussion where the user ends up wasting a lot of time and effort with no pay-off.

The ready-made patterns of Natural Conversation Framework offer even less help in trolling in debates than in the case of trolling virtual assistants, but they can help by providing a model for creating new intents and patterns for a given project. This also requires more research on how trolling works in debate situations. Building a debate trollbot based on conversational structures could be somewhat laborious, but the payback would be a troll bot that succeeds to cause harm in the context it is designed for.

A possibility of an all-purpose troll-bot?

One drawback in using the Natural Conversation Framework for creating a trollbot seems to be that the framework as well as trolling are both highly context-dependent. The endeavor requires knowledge on what kinds of projects, activities, and actions are typical for each kind of discussion, and each project needs a slightly different kind of trollbot. An all-purpose trollbot would need to have a lot of contextual knowledge as well as a huge range of intents and patterns it recognizes and can respond to. A somewhat more plausible option would be a platform-dependent trollbot. For example, one might assume that there are only a limited set of projects and activities in general forum discussions, which means that creating a troll bot for a specific forum would be feasible.

Table 3: A hypothetical discussion pattern between a user (U) and a trollbots (T).

Pattern		
01	U:	ARGUMENT
02	T:	DISCONFIRMATION
03	U:	ARGUMENT
04	T:	CHALLENGE
05	U:	QUESTION
06	T:	CHALLENGE
...		

Example		
01	U:	Something should be done to stop climate change fast
02	T:	Nothing should be done
03	U:	What do you mean? We will be in lots of trouble.
04	T:	You have no proof.
05	U:	Have you read the climate reports?
06	T:	You don’t need to know what I’ve seen. Where is the proof?
...		

CONCLUSION

We have hypothesized the possibility to create a successful troll bot – one that is capable of luring people into long, frustrating discussions – by utilizing concepts and ideas that arise from Conversation Analysis. A task like this would require a lot of work and good comprehension of how human interaction works, but the outcome might eventually surpass all trollbots designed thus far. Such a bot would enable more research into the phenomenon of trolling, as well as effective practices of preventing automated trolling.

ACKNOWLEDGMENTS

This work has been supported by Academy of Finland (grant nr. 320694).

REFERENCES

[1] Claire Hardaker. 2013. “Uh....not to Be Nitpicky„„ but...the Past Tense of Drag is Dragged, not Drug.” – An Overview of Trolling Strategies. *Journal of Language Aggression and Conflict* 1, 1 (2013), 58–86. https://doi.org/10.1075/jlac.1.1.04har

[2] Petra Jääskeläinen. 2020. *Conversation Analysis as a Design Research Method for Designing Socioculturally Contextual Conversational Agents*. Master’s thesis. Department of Informatics and Media, Uppsala University.

[3] Stephen Levinson. [n.d.]. Action Formation and Ascription. In *Handbook of Conversation Analysis*, Jack Sidnell and Tanya Stivers (Eds.). Blackwell Publishers, Chapter 6, 103–130.

[4] Robert J. Moore and Raphael Arar. 2019. *Conversational UX Design: A Practitioner’s Guide to the Natural Conversation Framework*. ACM, New York, NY. https://doi.org/10.1145/3304087