8-7-2015

Comparing Uncivil Discourse with a Conversational Intelligent Agent in Public and Private Online Settings

Mark Marcello

Comparing Uncivil Discourse with a Conversational Intelligent Agent in Public and Private Online Settings

by

Mark Marcello

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Human-Computer Interaction

Department of Information Sciences and Technology

B. Thomas Golisano College of Computing and Information Sciences

Rochester Institute of Technology
Rochester, NY
August 7, 2015

Abstract

According to a recent study by the Pew Research Center, 60% of participants of a survey said they have witnessed online abuse, and 66% of that abuse occurred via social media. This study looks at online communication, and tests if anonymity increases the amount of abuse that occurs during online communication. A quasi-experiment was developed to look for uncivil discourse using a conversational intelligent agent to create a controlled, repeatable conversational environment. The interactions with the agent occur in two settings: one in public at a kiosk, and a second in private. The researcher hypothesized that the interactions that occur in private would be more abusive than the interactions that occurred in private, due to the anonymity of the conversations.

Several thousand interactions were recorded in the two scenarios, and analysis was performed on the conversational data. A custom search engine used a data set of known abuse keywords to flag each conversation as abusive or not abusive. Using this abusive conversation count, a chi-square test of independence was used to determine that the number of abusive interactions with the agent was statistically higher in the private scenario compared to the public scenario. This finding supported the researcher's hypothesis, showing that users were almost twice as likely to use abusive keywords when interacting with the agent in a private setting.

Rochester Institute of Technology

B. Thomas Golisano College of Computing and Information Sciences

Master of Science in Human-Computer Interaction

Thesis Approval Form

Student Name: Mark Marcello

Thesis Title: Comparing Uncivil Discourse with a Conversational Intelligent Agent in Public and Private Online Settings

Thesis Committee

Name	Signature	Date

Michael Yacci
Chair

Anne Haake
Committee Member

Table of Contents

Background

Introduction

By definition, discourse is communication of thought by words, by use of speech or writing. It is not uncommon for individuals to express uncivil discourse, and the popularity of social media and other online-interactions appear to be increasing the amount of uncivil discourse readers are exposed to. According to a study by the International Business Times, "34 percent of Americans believe social media has had an overall negative effect on the quality of news and information on the Internet, compared with just 17 percent who believe it's had a positive effect" (Zara, 2012). This study will investigate online behavior, and look for relationships between online abuse and anonymity. The belief is that it is not the act of communicating online that promotes these negative conversations, but instead the assumed anonymity of the users that provides a sense of insulation from real-world consequences.

In this study, a controlled environment was created using an intelligent conversational agent to interact with a large testing group in both public and private settings. The agent allowed for repeatable responses and tone during interactions with users. The interactions were analyzed to discover if online communication in a private setting changes the behavior of a user compared to online communication in a public setting.

Online Harassment

Research in the area of online harassment is not new. According to the Pew Research Center, "It is a phenomenon that can take a variety of forms: name-calling, trolling, doxing, open and escalating threats, vicious sexist, racist, and homophobic rants, attempts to shame others, and

direct efforts to embarrass or humiliate people. While some accept online harassment as a nuisance, others face situations that prompt them to take serious action and precautions.
At a basic level, there is no clear legal definition of what constitutes 'online harassment.' Traditional notions of libel, slander, and threatening speech are sometimes hard to apply to the online environment. In addition, the anonymous and pseudonymous nature of the internet can make it easy for people to attack others without repercussions" (Online Harassment, 2014, pg 10).

The Pew Research Center (2014) surveyed 2,849 web users in June of 2014, and found that 60% of the participants said they witnessed someone being called offensive names, and 53% saw efforts to purposefully embarrass someone. They learned that of these online harassments, half of those harassed did not know who was behind it. The attacks were targeted against a specific person or group of persons behind the cloak of anonymity. Of those participants who witnessed some form of harassment, 66% of the harassments occurred via social media.

Kennedy & Taylor (2010) looked at online harassment and victimization on social networking websites from 354 colleges. The researchers found that certain types of harassment are more common in online settings, where other types of are more likely to occur in offline settings. "Some types of victimization occurred more in online settings, some more in offline settings and others in comparable rates online and offline. Two types of harassment occurred more frequently in online situations. First, sexual harassment occurred more frequently online. Second, pestering, incessant and unyielding attention also occurred more frequently in online situations. While being irritated to the point of wanting to end a friendship occurred in both settings, more students had their requests to cease contact ignored in online situations. The third type of harassing

behavior, verbal harassment, occurred in similar levels online and offline. The majority of students (53.4%) had been verbally attacked in one setting or another" (Kennedy and Taylor, 2010, pg 10).

Intelligent Agents

For this study, a technology solution was sought to create a controlled environment to test for this uncivil discourse. Reeves and Nass found that users tend to interact with technology as if the technology were human, even building relationships with technology. "Not only were the computers in these experiments tools for learning new information, they were social actors that people reacted to with the same polite treatment that they would give to another human. This certainly adds a new dimension to an understanding of human-media relationships." (Reeves and Nass, 1966, p. 32). Using this knowledge, it was inferred that a technology such as an intelligent agent would be a suitable method of creating a repeatable communication environment.

An intelligent agent combines natural language processing with a personality, often graphical, to interact with a user and host human-computer interaction. Different from most computer interfaces, agents have the ability to interact with the user in a close-to-human way, asking questions, problem solving, and even having the ability to look and act human. The idea of an intelligent agent taking the place of a human was first done in the 1960s, using the agent ELIZA, a well-known intelligent agent created by Joseph Weizenbaum. While multiple scripts, or response sets, could be used with ELIZA, the most popular one was DOCTOR, which played the role of a psychotherapist (Weizenbaum, 1966). Although giving the appearance of helpful conversation, ELIZA had a very limited set of responses, and had to rely on primitive natural

language processing algorithms to provide generic responses to questions or comments asked of it. Comments such as "My mother hates me" would be followed up by "Who else in your family hates you?" These responses would not be answers to questions, but would spark the user to solve the problems on their own, similar to a real-life psychotherapist.

Abusive Behavior Toward Agents

Research has shown that intelligent conversational agents can be the targets of abuse when used in educational settings. Kopp, Gesellensetter, Kramer, and Wachsmuth (2005) built a kiosk agent named Max, to be placed in a German museum to interact with museum visitors. Max collected and recorded information of the interactions between the agent and the museum visitors. "Visitors can give natural language input to the system using a keyboard, whereas Max will respond with a synthetic German voice and appropriate nonverbal behaviors like manual gestures, facial expressions, gaze, or locomotion. In doing so, he should be as natural and believable as possible a communication partner, being entertaining and fun to talk with" (Kopp, Gesellensetter, Kramer, and Wachsmuth, 2005 pgs 1-2). Max was projected on a large screen in the Heinz-Nixdorf-MuseumsForum, and logged the conversations with users over a period of time.

Kopp et al. found that the majority of users misused their kiosk. Abuse of the agent as well as testing the ability of the agent was discovered. Eleven percent of questions were flaming (abuse, name calling, senseless utterances), 3% asked questions about the agent itself, and less than 1% asked questions about the museum. Sixteen percent of the questions asked were testing the limits of the agent, asking questions such as "how's the weather" and "can you dance" (Kopp et al., 2005).

Banter with Bill is an informational kiosk agent and ongoing research project at Rochester Institute of Technology. Bill, based off of RIT President Bill Destler, uses a multi-round statistical analysis of input using TF-IDF to provide the best response to a question or comment (Yacci and Marcello, 2010). The audiences for Bill were attendees of the annual *Imagine RIT: Innovation and Creativity Festival* at RIT. Bill was specifically programmed to answer questions related to the festival, RIT, and questions about RIT President Bill Destler, yet was able to answer non-related questions to the best of his ability. Bill accepted user input via a keyboard located at the kiosk, and provided feedback through text displayed on a screen, as well as audible responses heard through speakers accompanied by lip-synched animation.

Abuse behaviors were discovered in an early study using Bill. The agent was set up and available as a kiosk agent for two years at the Imagine RIT festival. Data was collected during the festival's eight-hour period each year, and a content analysis of the recorded data was performed to determine categories of questions and comments. Categories titled "outsmart the system" and "misunderstanding the system" were determined to contain a large number of questions that fell outside the scope of our agent. The outsmarting category, although not as crude as some of the previous research showed, still contained abusive questions/comments such as "are you gay," and "you suck."

Doering, Veletsianos, and Yerasimou (2008) developed agents Penelope and Alex, that were used in a study of graduate students from a teaching degree program, asked to interact with agents as assistants to complete tasks using a software package called eFolio. Although the agents were programmed with program-specific content, they were also conversational in nature,

and were able to answer personal and general questions asked of them by the participants in the study. Users grew attached to their agents over the 4-week study, and researchers discovered some key information for designing successful agents. “If learners are to utilize conversational agents with success, the agents need to be intelligent enough to accurately comprehend the questions the learners are posing or the software application needs to offer an interface that guides the learners to ask their questions in an appropriate manner” (Doering, Veletsianos, and Yerasimou, 2008, p. 267). Although 70% of the participants enjoyed the agent and believed it was helpful in providing assistance during the learning process, many were frustrated when the agent failed to provide correct answers on seemingly easy questions. The researchers found both male and female participants asked sexually explicit questions. One participant’s reasoning for this behavior was, “Because she had an answer for everything, it was like pushing her limits. You know it wasn’t a real person, so it’s not that you are offending her” ” (Doering et al., 2008, p. 264).

In another study, Veletsianos, Scharber, and Doering (2008) used an agent in a middle school classroom exercise with participants ranging from 14 to 15 years of age. Students interacting with this female agent asked a wide variety of sexually explicit questions, while also becoming extremely abusive toward the agent, with comments such as “stop looking at me,” and “you are a whore” (Veletsianos, Scharber, and Doering, 2008, p. 298).

The studies that report uncivil communication and agent abuse are not experiments; there is no control group. Hence we know that abuse and uncivil behavior occurs online but no experimental studies were found that controlled for the *public and private nature* of online communication.

The researchers theorize that the private anonymous nature of social media and other internet-based communications are what cause a hostile environment, not the act of using social media by itself. The belief is that if users' online communication occurred in public, they would be more likely to be polite due to the accountability of having other users present watching them interact. Others would witness any abusive comments they would have with the agent. In an opposite manner, users who interact privately would have no one watching them, so they could be more "raw" with their conversation.

Hypothesis

The null hypothesis states that the amount of abuse observed in a private online setting would be less than or equal to the abuse observed in a public online setting. The alternative hypothesis states that the amount of abusive conversation recorded in private will be greater than the abusive conversation recorded in public.

Methods

To test this hypothesis, a quasi-experiment was created using two online scenarios to be used as research treatments. In both scenarios an intelligent agent was used to simulate online communication. The agent conversation is "pre-programmed" thus controlling for conversational factors such as tone of voice, intentional offending someone, or other conversational actions that might inflame a conversation. The agent was programmed to deliver relatively innocuous conversation, acting as a constant in the study. The treatments varied according to the setting: interacting with the intelligent agent in public at a kiosk surrounded by other people (public interaction), and interacting with the agent in online via a website (private interaction).

Conversational data from both of these scenarios was stored by the researcher for later analysis, making it possible to determine the amount of certain conversational actions. Of interest to our study is the amount of abuse that appears in comparison to helpful conversation. The data will allow the analysis of the results to determine if there is a greater amount of abuse in the private interactions compared to the public interactions. Details on the development of the conversational agent are included in Appendix A.

The study was performed in two parts. The public interactions were gathered over a 2-day public exhibition that was held at the USA Science and Engineering Festival in Washington, D.C. The private interactions were gathered in the following two weeks from a website. The in-person situation allowed public interaction with an agent. The online situation used the same agent, but the conversations were not public. The agent took on the persona of Albert Einstein, and answered questions in character.

Public interactions were gathered as the intelligent agent was projected on a screen. Users could ask a question of the agent using keyboard and text that was also displayed on the screen. The agent response was both text (again, projected on the screen) and also auditory, via attached speakers. All interactions were plainly visible and public to the participants who were watching the exhibit. Private interactions were gathered as the same agent was posted on a website that allowed a user to ask questions of the agent online. However, in the private, online scenario, the user could not see any other interactions – hence the private nature of the conversation.

The measure of abuse in this study was done with a post-categorization of the data, using a custom search engine that processed the results logged by the systems, searching against a known list of abusive words. These words were collected from an online resource that provides abusive word lists for spam filters and other natural language processing needs (Bad words list, 2008).

A PHP script was written to perform this search following this general algorithm:

1. Search program reads in the entire array of user comments (A).
2. Search program reads in the entire array of abusive keywords (B).
3. While looping through array A, each iteration loops through array B, looking for a match.
4. If an abusive keyword is found in the user comment, a variable (C) is incremented, and skips to the next iteration of array A, until complete.
5. The variable C is used as the measure of abuse within the set of user comments.

By implementing this type of search, and using the known set of abusive keywords, it was possible to eliminate human error in counting, as well as bias in deciding what constitutes a keyword. This script would be able to be run on any text file-based list of comments that are new-line delimited. This type of key word counting cannot catch subtler forms of harassment that contains elements of sarcasm, so it was accurate and unbiased as a measure of abuse, although not complete.

Additional Details of the Public Scenario (Washington, DC)

The booth was set up in the National Mall in Washington, D.C., as part of the USA Science and Engineering Festival. Over the course of a 2-day period, the agent, Ask Albert, was set up in an 8' by 8' outdoor booth that was in heavy traffic from the festival. A constant flow of users were at the exhibit for 12 hours both Saturday and Sunday. The agent ran continuously throughout this time, logging a total of over 1,488 interactions.

The computer hosting the agent was set up with a simple user interface, allowing users of all ages to interact with Ask Albert. A 6' table was set up with a computer screen facing the users, pre-loaded with the Ask Albert software. The users were given a keyboard with the curser already set up in the text box, so that they could type anything they wanted; once they hit enter, Albert would display the response in a text bubble on the screen and speak the answer through a set of desktop speakers located at either side of the monitor.

Participants were not coached in any way on what to ask the agent, but were given some examples of questions when they did not fully understand the exhibit. For example, if a user did not understand, the researcher would provide a prompt, such as, "Ask Albert can answer questions about Albert Einstein's life, about science and engineering in general, as well as questions about the festival here. You can also ask him random questions, and see how he will respond." If the user still did not understand the exhibit, a question was provided for them, such as "How old are you?"

Experiences from the Booth

Due to the type of festival, participant's age ranged from young children to adults. Most of the participants were comfortable using the computer, while a small percentage needed help, which could include typing for them or explaining what the agent's response was if they could not see or hear it properly. Anecdotally, throughout both days, many users would leave with smiles on their faces, finding the agent either fun to use, or enjoying the response of the agent, whether it was an accurate response or not. This observation was also witnessed with our past studies; where in early stages of the agent it was very inaccurate, users still found it entertaining.

Not all users with our kiosk left happy though. One user was very upset with the system because of its response to a question she asked. One of the categories of questions the system could answer was RIT and Rochester, a category carried over from past iterations of the agent where it was used in Rochester N.Y., at the *Imagine RIT: Innovation and Creativity Festival*. The system is loaded with questions and answers about Rochester Institute of Technology, important people at the university such as RIT President Destler, and general information about Rochester N.Y. This specific user was upset about the answer to the question "what is your favorite color?" The agent responded, "Orange is my favorite color. RIT's school color is orange." After the agent responded, the user asked if orange was indeed Albert Einstein's favorite color. The researchers explained that Einstein's favorite color was unknown, but the agent's favorite color is orange. The participant then left the booth, upset. While this interaction was intended to be playful, apparently, the user was using Ask Albert for informational purposes. While humans can convey a playful response with facial movements and tone of voice, Ask Albert had neither of these capabilities.

Details of the Private Scenario (Internet Study)

The private scenario occurred via a website and worked functionally the same as the booth above, but users were able to ask the agent questions in the privacy of wherever they were accessing the website. The users for this part of the experiment were given a 2-week period to interact with the agent via the website. Users were not selected by the researchers, but instead were asked to participate through multiple venues, including Facebook and multiple online forums, as well as through the documentation handed out the day of the festival in D.C. The age range of users for this part of the study was not known, as no personal information was collected from the system. 1,009 recorded responses were collected for analysis over this 2-week period.

Results

This study was concerned with counting the number of abusive occurring in the conversations and interactions with the agent. It should be noted that the measure of abuse in this study is limited to the set of abusive keywords that were counted in the conversations. As described previously, this is a commonly used text corpus but it is not capable of detecting subtlety in inflection or inference.

The abusive keyword search was used to count the interactions. The private interactions with the agent produced 40 abusive interactions out of 1011, while the public, festival-based interactions with the agent produced 37 abusive interactions out of 1488.

A Chi-square test of independence was calculated comparing the frequency of abusive comments in public and private settings. A significant interaction was found (chi-sq = 4.356, p = .0369). In

this study, there are significantly more abusive statements made in private than in public. This allows rejection of the null hypothesis that the amount of abuse recorded online is equal to the amount of abuse recorded in-person at the kiosk.

The Chi-square test is commonly used to determine there is a relationship between the row variables and the column variables. A probability of this relationship is reported in a p-value. The common threshold of $p<.05$ is used in this study to mark the region of rejection in the test. To calculate this, tables are generated of expected and observed values from the study for use in the Chi Square statistic. The observed count of abusive and non-abusive interactions are shown below in Table 1.

	Private/Online	**Public/Kiosk**	**Totals:**
Abusive	40	37	77
Non-abusive	971	1451	2422
Totals:	1011	1488	2499

Table 1: Observed Count of Interactions

The expected values can be calculated for each cell by taking the total of a column and multiplying it by the total of a row, and then dividing by the grand total. For example, using the data above, to calculate the expected value of abusive conversation online: 1011 * 77 / 2499 = 31.15. Performing this process for each relative cell gives the expected value for each cell. As can be seen in Table 2 below, the expected values differ from the actual values.

	Private/Online	Public/Kiosk
Abusive	31.15	45.85
Non-abusive	979.85	1442.15

Table 2: Expected Values

Using these two tables, the Chi-sq statistic of 4.356 is calculated, resulting in a p-value of 0.0369, allowing us to reject the null hypothesis. These findings support the alternative hypothesis stating that interactions in private lead to more abusive behavior than interactions with in a public setting.

Conclusion

Based on a number of factors, including past studies with intelligent agent abuse, the prevalence of online bullying in today's social media, as well as observed behavior of people when speaking online with anonymity, it was believed that in this study using an intelligent agent; the private interactions would be more abusive in nature than the public interactions. This hypothesis was supported by the results of this study, showing that users were almost twice as likely to use known abusive keywords when interacting with the agent in a private setting.

The keyword search that was used in this study identified only 2.49% of the public interactions being abusive and 3.69% of the private interactions being abusive. The numbers of abusive questions were overall lower than expected, based on the researcher's anecdotal observation of users. One possible explanation for this low percentage is that we were only able to search for

known abusive keywords, and didn't look at other possible misuse of the system or more subtle insulting or harassing conversation.

Future Studies

Given this knowledge that anonymity appears to be the main cause of online harassment, future studies should attempt to find ways to mitigate this behavior, presumably by decreasing the amount of anonymity we use online. Facebook requires user accounts to be tied to real people, and the company often removes fake accounts from their system, where as Twitter, Reddit, and other social media systems do not require this level of accountability. Other possible studies could look into how much this uncivil discourse that occurs online in anonymous situations affects offline communication. Reducing online harassment by itself may not seem too important, but if this harassment is tied to depression, real-life abuse, and suicide, it could be a very meaningful and powerful area of study.

Acknowledgements

I would like to thank my advisor for many years of encouragement, mentoring, and education. Without Dr. Yacci's help and knowledge, this study, as well as all of the ones leading up to it, would have never been able to happen. I would also like to thank Rochester Institute of Technology for sending me to Washington, D.C., to participate in the USA Science and Engineering Festival, as well as allowing me to showcase my agent exhibit at the festival.

References

Bad words list. (2008, April 4). Retrieved January 2, 2010, from http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/

Doering, A., Veletsianos, G., & Theano, T. (2008). Conversational Agents and Their Longitudinal Affordances on Communication and Interaction. *Interactive Learning Research, 19*, 251-270.

Gustafson, J., Lundeberg, M., & Liljencrants, J. (1999). Experiences from the development of August – a multi-modal spoken dialogue system.

Kennedy, A., & Taylor, M. (2010). Online Harassment and Victimization of College Students. *Justice Policy Journal*.

Kopp, S., Gesellensetter, L., Kramer, N., & Wachsmuth, I. (2005). A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application.

Online Harassment. (2014, October 22). Retrieved June 26, 2015, from http://www.pewinternet.org/2014/10/22/online-harassment/

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Stanford, Calif.: CSLI Publications ;.

Veletsianos, G., Scharber, C., & Doering, A. (2008). When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers, 20*, 292-301.

Weizenbaum, J. (1966). ELIZA-A Computer Program Fore the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM,* 36-45.

Yacci, M., & Marcello, M. (2010). Questions and Response Accuracy in a Kiosk Agent Context. *Autonomous Agents and Multiagent Systems*.

Zara, C. (2012, December 1). Is Twitter Making Us Meaner? Uncivil Discourse In The Age Of Social Media. Retrieved June 26, 2015, from http://www.ibtimes.com/twitter-making-us-meaner-uncivil-discourse-age-social-media-909856

Appendix: The History of the Test Agent

The agent used for this study was actually the third version of our agent software that has been developed over the past 5 years. The first version of the agent, Banter with Bill v.1, was created for the second annual *Imagine RIT: Innovation and Creativity Festival*. Banter with Bill v.1 was a very simple agent that was built from existing code that advisor Michael Yacci had used in the past for a course he taught involving conversational agents. The system was built using JavaScript and XML to parse the users' input and find the best answer, and then used Microsoft Agent to actually move the eyes and mouth of our animated head on the screen as well as speak the response to the user. This first version of our agent gave us good ideas of how random users would interact with the agent, and set the groundwork for building future agents with both improved backend coding, and improved conversational accuracy.

The second version of our agent, Banter with Bill v.2, was created for the third annual festival at RIT. In this build of the agent, we moved away from the XML and JavaScript system, which was a very brute-force approach of natural language, and instead implemented a two-step process of the statistical measure term frequency-inverse document frequency. TF/IDF is an efficient and simple algorithm used in classifying large amounts of data. While there are other algorithms available for matching words in a query to a dictionary of content, TF/IDF is lightweight and simple to use, making it a good fit for an algorithm to test with our agent.

The goal of Bill V.2 was to use TF/IDF in real time as part of a two-stage analysis of user input. The first stage of analysis was to best determine the category in which the user comment belonged. These categories were pre-designated ahead of time, based on the content analysis

results of previous years' studies. These categories consisted of topics such as festival, RIT and Rochester, about Bill, personal/general, misunderstanding the system, and abuse. Our theory at the time was that if we were able to first categorize the interaction, we would be able to increase accuracy of the response by narrowing down the pool of potential responses by 90%. The second round of classification occurred within that specific category, and netted the correct answer rated by the highest probability. This system worked as intended, but its relatively low accuracy of 65% correct classifications showed there was much room for improvement. The biggest problem we found was that if comments were being wrongfully classified in the first phase, there would be a 0% chance for the correct answer to be selected in the second round. Therefore, it was essential for the first phase to be near perfect in its accuracy.

The third version of the agent, Ask Albert, was built from the existing TF/IDF based code used in Banter with Bill v.2, but was altered to fit the needs of this study. First, the agent was changed from the likeness of RIT President Bill Destler to a more recognizable figure in history, Albert Einstein. A dataset of responses and categories was tailored to the USA Science and Engineering Festival, including categories of conversation such as Washington, D.C., Albert Einstein, science, engineering, physics, etc. The same basic two-stage approach was used as mentioned above in Banter with Bill v.2, but updates to the code to improve accuracy and speed of the system were performed in the year leading up to this study. We projected an increased accuracy of response from the previous year's 65%, and we achieved that with a nearly 72% accuracy rate combined from the two trials explained below.