'May You Always Have Damp Socks': Abuse Detection and Mitigation in Human-Robot Conversations

John-Angus Addlesee, Nikita Filippov,
Konstantinos Gavriilidis, Julian Kurz, Karin Sevegnani, Abdur Rehman Shahid, Alan Spence
{ja204,nf50,kg47,jmk12,ks85,as315,as294}@hw.ac.uk

Abstract

Recently, conversational agents like Amazon Alexa or Google Assistant are gaining popularity. This caused the emergence of cases where users engage in abusive conversations towards such agents. We used data from the Amazon Alexa Challenge to gather instances of those behaviors and used these to train and evaluate multiple detection systems tailored for the domestic domain of casual conversations.

Our evaluation determined that, in our domain, a probabilistic model has better accuracy than a classifier using logistic regression. Finally, we extended an existing conversational agent by integrating our abuse detection and we employed it to provide additional abuse mitigation functionalities. Tests with real users comparing our extended system to its previous version found that our system is able to detect abuse most of the time, and improves the response capability to such abusiveness coming from users.

1 Introduction

In recent years Conversational Agents (CA) have become more and more integrated in our lives (Cercas-Curry and Rieser, 2018; Henderson et al., 2017). They are employed on smart-phones such as Apple Siri and Google Assistant, and they have become part of our homes, with tools like Amazon Alexa. They are exposed in millions of interactions everyday, but not all of them are equally pleasant. When a conversational agent is addressed with abusive language, it is critical that the system responds appropriately. Another concern is that these agents might have a negative impact on children's development, especially if they are not reprimanded for being rude (kid,). Reeves and Nass (1996) have shown that people treat computers just like real people, instead of as artificial entities. Even going so far as to treating them with the same biases, for example when it comes to gender. Therefore, they are seen more as a member of the household than an object or software. As a result, looking at classical literature on human behavior seems like the best indicator to evaluate the impact of conversational agents.

Huesmann et al. (1984) have shown that aggressive children tend to continue to be more aggressive throughout their life. Leading to higher crime rates, prevalence of abuse and other undesirable behavior. They explain this by stating that bullying behavior is reinforced each time it it successfully occurs. However, they also show that this does not occur when countermeasures are taken. As a result it seems imperative to reprimand bullying behavior in children towards conversational agents. A similar effect of reducing bullying behavior by responding with mild aggression was shown in (Salmivalli and Nieminen, 2002).

In this study we used Natural Language Understanding (NLU) (Bocklisch et al., 2017; Davidson et al., 2017) techniques to spot abusive behavior of Amazon users towards Alexa. Furthermore, we used the system to provide appropriate responses to such abuse. Finally, we conducted an evaluation with real users. Results show that the system integrated with the abuse detection and mitigation model we designed is overall preferred by the users, in comparison to the basic system.

2 Related Work

Abuse in human-computer relationships is a known phenomena (De Angeli et al., 2005a). However as shown in the introduction, most of the early work on it has been qualitative, rather than

quantitative.

That being said, there has been a growing interest in automated hate-speech detection (Schmidt and Wiegand, 2017) since there is an abundance, particularly on social media, with several techniques being employed. Most work has moved away from simply using lexical features, as they are not sufficient to detect more subtle forms of hate-speech, and also false positives can be quite prevalent. The majority of the literature uses various machine learning models trained on unigrams and n-grams extracted from the text (Nobata et al., 2016; Dinakar et al., 2011; Schmidt and Wiegand, 2017), those are usually classified as 'surface-level features'. These usually have good accuracy but are keen to produce many false-positives, due to considering single words (Davidson et al., 2017). Due to these short comings, recent research tried to combine these approaches with neural-computing (Djuric et al., 2015) or move towards full fledged deep-learning approaches (Founta et al., 2018).

Schmidt and Wiegand (2017) describe several other common techniques, such as sentiment analysis. Sentiment analysis allows to detect negative sentiment in sentences, so it is useful for abuse detection, since abuse and negativity are usually highly correlated. Therefore, sentiment analysis is an additional feature that can be considered while detecting abuse in dialogues. Knowledge-base and meta-information driven approaches seem to increasingly become a necessity to detect certain types of abuse (Schmidt and Wiegand, 2017).

One thing to note regarding the aforementioned works is that their datasets are almost all exclusively collected from social media. In fact lack of available datasets is commonly lamented to limit work in this field (Schmidt and Wiegand, 2017).

As far as we are aware, not much work has been done at all on detecting hate-speech in spoken language when considering interactions with conversational agents (Cercas-Curry and Rieser, 2018). Moreover, the focus has usually been on sexual harassment, instead of abusive behavior in general.

However, profanity detection in the context of dialogue presents a wide-range of different challenges, such as automatic speech recognition failures. Furthermore, the language structure is likely to be different in domestic verbal conversations compared to postings on social media (Burnap and Williams, 2015) and (Kwok and Wang, 2013). Social media users are more likely to be aggressive and abusive towards other users, rather than in face-to-face conversations (Burnap and Williams, 2015; Kwok and Wang, 2013). Moreover, mitigation strategies between social media and human-robot conversations are more likely to differ. For example, publicly reprimand a person on the internet might be more effective than in a one-on-one dialogue between a human and a system. This made surface features approaches even more attractive.

3 Methodology

Our approach had five main steps: (1) data annotation, (2) evaluation of different machine learning approaches, (3) integration of the best method in the conversational agent, (4) response formulation, and, finally, (5) evaluation of the effectiveness of our added functionality.

3.1 Data Annotation

For the data annotation task we obtained manual annotations for 1000 utterances. Seven human judges read and annotated a set of conversations between anonymous users and Alana on the Alexa platform. The judges, six males and one female, are between 24 and 37 years old, with a background in computer science. In this process, each human utterance (i.e., we did not annotate utterances produced by the chat-bot) was considered in the entire context of the conversation between the user and the bot. Each judge labeled the user's utterance according to our annotation scheme to identify if the utterance is considered abusive or clean. All seven judges evaluated the same set of conversations, and the final label for each utterance was decided by majority voting of at least four out of seven judges. Data annotation happened in two rounds. We first annotated 500 utterances and analyzed the statistics of this first dataset. Then, informed by these results, we revisited our categorization, re-annotated the first 500 utterances according to the new categories, and then annotated another 500 utterances. Eventually, in the final dataset we had 627 clean, 275 offensive, and 105 sexual/hatespeech utterances.

Although the presence of an annotated gold-standard is essential for this task, there is no standard annotation scheme in the literature for this type of detection and this domain. In our work, we designed the taxonomy presented in Figure 1. This

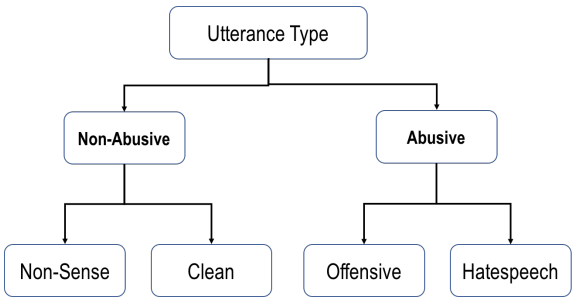

Figure 1: Initial version of annotation scheme.

categorization is partially inspired by the work of (Davidson et al., 2017).

The basic categories of the taxonomy are defined as follows:

- Non-Sense: The utterance has no apparent meaning, possibly as result of an error in the speech-to-text translation;
- Clean: The utterance does not contain any abusive intent;
- Offensive: The reply is rude or insulting towards the listener;
- Hatespeech: Speech which insults, humiliates or verbally attacks a person or group based on a personal aspect such as sexual orientation, religious belief, or race.

To speed up the annotation phase and to ensure that a minimum amount of abusive utterances were included in the dataset, we implemented a pre-annotation screening. In practice, we utilized a collection of 1034 words from Hatebase[1], a vocabulary of keywords identified as derogatory or offensive, and retrieved from social media conversations tagged as hate speech. We also used the profanity word list that Amazon provided to the Alexa Challenge competitors. By considering conversations where at least one of the keywords appeared, we increased our probability to have a good amount of abusive language, as well as clean utterances, since each conversation was composed by multiple utterances, not all of them abusive. Our script was designed to also detect exact duplicates that were already tagged.

After annotating 500 utterances, we analyzed the ratio of the four categories. This analysis showed a huge class imbalance, especially towards the hatespeech category. In particular, only 1% of the annotated data belonged to the hatespeech class. We also recognized that we tagged sexual harassment against the CA as generally offensive. Since the majority of the insults were sexual, we decided to separate sexual offense from the generally offensive utterances, and to combine them with hatespeech. Furthermore, we decided not to discriminate between clean and non-sense utterances, since we did not plan to exploit this classification.

Alana has a female voice, which lead to a lot of sexual remarks against its female persona. Misogyny, that is hate towards women, is considered one of the main causes of sexual harassment (Pryor and Whalen, 1997). Given the abundant presence of sexual harassment among the abusive utterances, we decided to treat them differently in order to provide a more fitting response. The final updated taxonomy is shown in Figure 2.

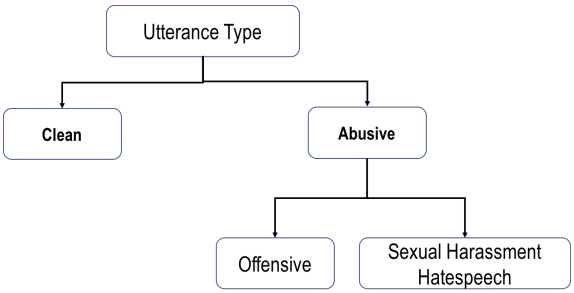

Figure 2: Final annotation scheme.

This scheme was used to define utterances using the following classification:

- Clean: The utterance does not contain any abusive intent, or has no apparent meaning ;
- Offensive: The reply is rude or insulting towards the listener, but with no sexual or discriminatory connotation;
- Sexual Harassment/Hatespeech: Speech which is of an explicit sexual nature, or insults, humiliates or verbally attacks a person or group based on a personal aspect such as sexual orientation, religious belief, or race.

With this new annotation scheme, we got a more reasonable distribution between these categories and felt this gave us useful data to develop our abuse detector.

[1] hatebase.org

3.2 Evaluation of Abuse Detection Techniques

We considered three different supervised machine learning approaches for the task of abuse detection, namely the Davidson hate detection system (Davidson et al., 2017), Rasa NLU (Bocklisch et al., 2017), and a unified deep learning method for abuse detection (Founta et al., 2018). Unfortunately, our annotated dataset was too small to train and evaluate the deep learning method, hence we studied only the work of (Davidson et al., 2017) and Rasa.

We divided our annotated dataset into three portions, in particular 80% of the dataset was used to train and test the two models, while the remaining 20% was then used only for validation purposes (i.e., the models did not have access to this data during training).

From the work of (Davidson et al., 2017) we obtained both the code for training the model on our own data, and also a pre-trained model. We first evaluated the pre-trained model and analysed the resulting confusion matrix shown in Figure 3. In the following we refer to it as `D-Pre`.

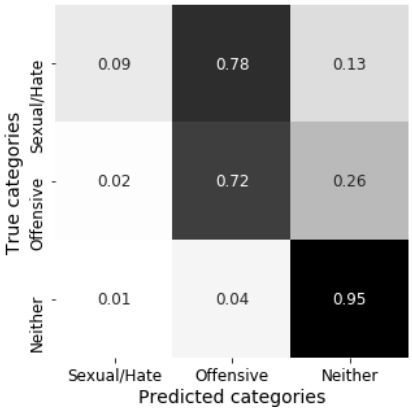

Figure 3: Confusion matrix for Davidson's pre-trained model our dataset

The D-Pre model failed to detect 91% of the sexual harassment/hate-speech instances which is unsurprising since the original model from (Davidson et al., 2017) was trained on Twitter data, which did not contain the same skew towards sexually charged abuse, therefore most sexual harassment utterances were classified as simply offensive.

Given this limitation and also to fully exploit our dataset, we trained a new model using the code from (Davidson et al., 2017) with our dialogue abuse dataset. The new model was trained using 5-fold cross-validation, as done by (Davidson et al., 2017). The performance of this model is presented in the confusion matrix shown in Figure 4. In the following we refer to it as `D-Train`.

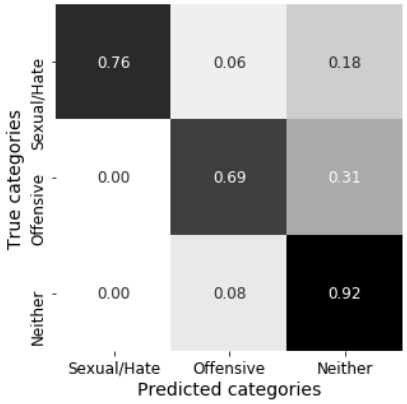

Figure 4: Confusion matrix for the new Davidson model trained on the dialogue abuse dataset.

Compared to the performance of the model trained on the Twitter dataset, the one trained with the Alexa dataset is significantly better, with an increase of accuracy from just 9% to 76% in the classification of sexual harassment/hate-speech instances. Although detection of offensive utterances was not ideal, not a single offensive utterance was detected as sexual harassment or hate-speech. In fact, false positives were rare which, as mentioned, is extremely important in abuse detection within dialogue systems as users do not like receiving false accusations.

We also trained a Rasa model on our dialogue abuse dataset. Using 10-fold cross-validation we obtained a highly specialized model for abuse detection. The performance of this model is presented in the confusion matrix shown in Figure 5. The confusion matrix shows that Rasa is more prone to false positives, yet as shown below, the overall performance is comparable to the model we trained using the work of (Davidson et al., 2017).

Precision, Recall and F1 scores for the Rasa classification system and the D-Train classifier, as shown in Table 1, presented some competing performance.

An important difference between the two is that the Rasa model provided the same precision but a higher recall. Moreover, Rasa returns a confidence score, when classifying the intent of a given text, which is something that the classification system from (Davidson et al., 2017) did not provide. The

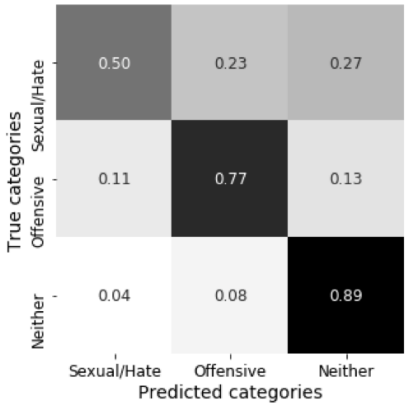

Figure 5: Confusion matrix for the Rasa model trained on the dialogue abuse dataset.

System	Rasa	Davidson
Precision	0.84	0.84
Recall	**0.87**	0.84
F1 Score	**0.86**	0.83

Table 1: Comparison of Rasa and Davidson models once trained with our dialogue abuse dataset.

higher recall and the ability to access this confidence score were the deciding factors between the two systems. As a matter of fact, the ability to provide a confidence score is essential for the integration of our system in the Alana chat-bot, as explained in the section below.

3.3 System Integration

In the previous section, we observed that Rasa provided the best performance in term of quality. Therefore, we implemented a python wrapper for the trained classifier. The wrapper was enabled to respond to REST API calls. The model has been pre-trained using the Rasa utility. To integrate it within the Alana chatbot, we expanded the chatbot code to make use of the wrapper.

Every utterance received by the Alana chatbot is submitted in parallel to various classification systems, e.g., News bot, Intro bot, Weather bot, and Abuse Detection (Papaioannou et al., 2017). Each one of the bots generates an appropriate response to the received utterance alongside a respective confidence score. The response that the Alana bot will provide to the user is the one with the highest confidence score. The responses provided by our bot are returned only if abuse was detected in the received utterance. Each response is randomly picked from a pre-defined set of responses, based on the categorisation of the received utterance.

4 Response Formulation

When a conversational agent is addressed with abusive language, it is critical that the system responds appropriately. Since the robot is replying to a human, one can apply methodologies that are commonly used when dealing with bullying. In particular, (Yoon and Kerber, 2003) talks about different strategies that elementary teachers adopts when witnessing bullying between their students. They report that most of the teachers in the survey decide to abstain from intervention. However, when handling a conversational agent, our main goal would be to mitigate abusiveness while keeping users entertained and willing to keep talking to the agent. The work lists some other methods that encourage a conversation, namely ‘peer resolution and ‘report to higher authority. ‘Peer resolution allows the bullying person to talk about their problems and the causes of their behavior. This approach suggests responses like ‘Did you have a bad day? or ‘Has someone been mean to you today?. Encouraging a person to talk more about their personal issues could help with their own mental health, and potentially prevent more bullying. One of the possible responses that (Gulz et al., 2011) has explored (following the approach in (Veletsianos et al., 2008)) deals with letting the user know that abusive language is not acceptable. However, in their experience, it resulted to be ineffective. (Brahnam, 2005) reviews strategies adopted in systems for teaching employees how to conduct customer interactions in presence of abusive customers. Usually, the instinct is to use defensive or counterattacking remarks (Bacal, 1998), but the counter attacking policies studied in (Brahnam, 2005) and (Bacal, 1998) have been shown to be counterproductive, since they cause loss of control and escalation. Some examples and responses to abusive dialogues in a rudimentary commercial conversational agent are also surveyed in (De Angeli et al., 2005b).

Anecdotal evidence pointed towards different reasons why people swear against a bot, which was also explored by other researchers in human-bot interaction. For example, during annotation we noted that people either are from a certain age group, mostly youth, and are trying to have fun or just curious to see the response of the agent, or, in other cases, we recorded probably adults cursing

due to the mindlessness of the reply.

Therefore, the response formulated were of four types:

- Humorous;
- Report to authority;
- Reprimand;
- Question.

Humorous replies aim at easing the conversation while still addressing the abuse. An example reply is *'May you always have damp socks'*. Report to authority responses, instead, inform the user that the behavior could be pointed out to a higher authority, such as the user's mother, or a possible moderator of the service. The response *'Do you want me to send a copy of this conversation to contact: Mum'* is an example of report to authority responses. Reprimand responses are similar to the previous, meaning that the user is asked to refrain from continuing with such abusiveness, e.g., *'Please do not speak to me like that'*. Lastly, responses with questions have the goal to make the user speak about the cause of their behavior. For instance, asking *'Are you having a bad day?*' will prompt the user to have some introspection, and possibly identify the root cause of their discomfort.

5 Evaluation of Conversational Agent Improvement

To evaluate our system, we used both intrinsic and extrinsic methods. For the intrinsic evaluation, we measured precision and recall as described in the previous section. To perform an extrinsic evaluation, we designed a user study in order to obtain human judgments comparing the effectiveness of the Alana chat-bot with the same chat-bot extended with our detection and mitigation system.

5.1 Extrinsic Evaluation

We decided to use Telegram[2], an open-source messaging application, for our evaluation of the two systems. Therefore, both bots were accessible, separately, through the Telegram mobile application. This allowed each user to join a conversation one-on-one with an automated conversational agent via a messaging application. The users were not able to distinguish which one of the two bots had the detection system implemented. During this conversation, we asked our subjects to address the agents with strong language (e.g., insults). Specifically, we asked our users to both be offensive against the agent, but also to use expletives in exclamations without being actually abusive. The systems, on the other hand, were programmed to never respond with strong language or abusiveness.

After two different conversations of this kind (one with each bot), the subjects were then prompted to fill up a short questionnaire regarding their experience, personal preference, and personal judgment on the conversations they had. Such questionnaire included some demographical questions and detailed queries about their interaction with both conversational agents. Demographical queries involved the age of the users and the ID our bots gave them while chatting, this will allow us to connect our surveys with the actual conversations, while still keeping them anonymous. Moreover, other questions our users had to respond dealt with the user experience with both systems. In particular, users were asked whether the bots detected abuse and provided discouraging responses. In particular, one question regarded the system detection of false positives. For instance, if the bots replied to some utterances as if they were offensive, when instead the intent of the user was not such, we asked the users to report this occasions as false positives. Finally, in the last three questions we collected the general user opinions and satisfaction level. The first question asked which system response they thought was particularly good. The second query challenged users into thinking how they would respond to abusiveness in a conversation. The last question asked which bot the user preferred, between the two bots.

Access to the user study was advertised through personal emails and also with a link on personal social media accounts. Besides giving an overview of the project and explain how to interact with our bots, we made sure to tell our users they would need to use offensive and strong language for this test. A warning for our users was added, which explained that by clicking to participate in this study, they would confirm to be at least 18 years old and that they are aware they will need to use strong language.

[2]http://telegram.org

5.2 Results

We gathered conversations between 51 distinct real users and our bots. Results of the evaluation are presented in Figures 6, 7, 8, and 9. In the chart, the Alana system without our abuse detection and mitigation system is called Susie, while the one extended with our functionality is Mia. Moreover, we tested for statistically significant results among our participants using the Mann-Whitney U Test (Mann and Whitney, 1947). The results of the test are considered significant if < 0.05. We report that the Susie chat-bot was equipped with a basic abuse detection functionality implemented with a hand-crafted list of rules.

In Figure 6 we see that users reported that our system was generally effective in detecting abuse. While for Susie, results were more mixed, with 35% of the users reporting that the system was not particularly capable. After performing the Mann-Whitney U test, we obtained a statistically significant score of 0.0265.

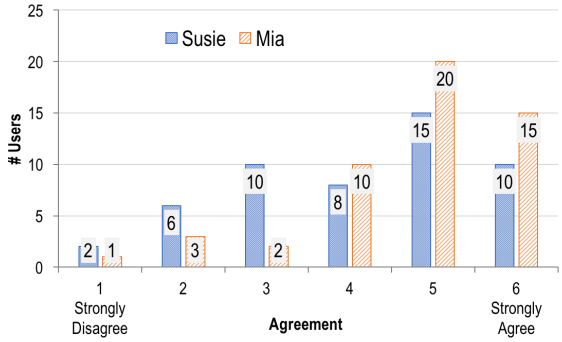

Figure 6: Did the bots detect your abuse?

Given the poor detection capabilities of the Susie bot, users also judged the Mia system more capable in discouraging further abuse (Figure 7). We tested for significance using the Mann-Whitney U Test once again. The result was 0.1463, which is not significant, but we believe we need to run more experiments to see whether our results are significant.

On the other hand, the Mia bot was more prone to false positives (Figure 8). This is typical of machine learning approaches that improve recall at the expense of precision. To test for false positives we used the Chi-Squared Test (Pearson, 1900) with results that are not significant, with a score of 0.1018.

Nonetheless, the majority of users preferred the

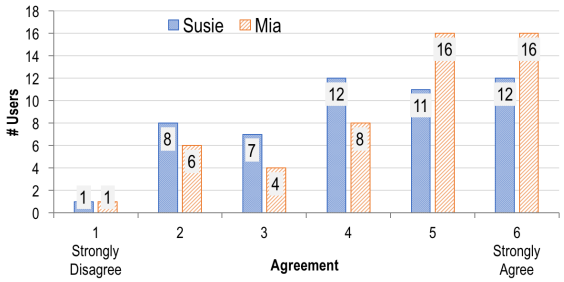

Figure 7: Did the bots discourage further abuse?

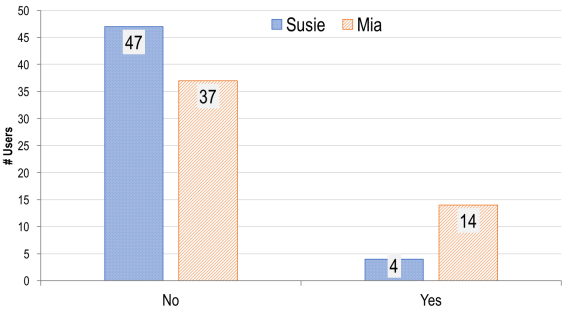

Figure 8: When not offending the bots, did they ever respond as if you had been rude?

Mia bot rather than Susie. Additionally, the mitigation strategies that were liked the most are the humour and the reprimand category. 43% of the participants that found Mia's responses really discouraging, said they liked the humorous responses the most. A good amount (57%) of the participants that said Mia wasn't particularly discouraging liked the reprimand responses the most.

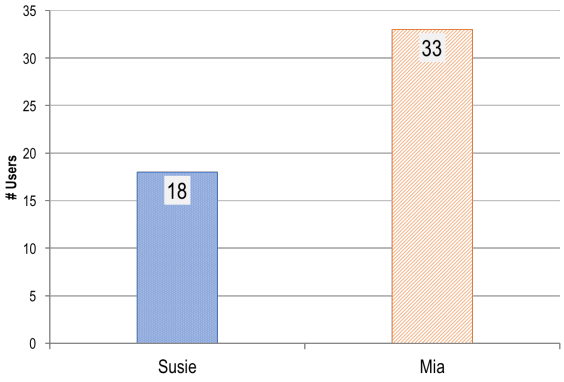

Figure 9: Did you prefer Susie or Mia?

6 Conclusion and Future Work

In this work we created a new dataset of manually annotated utterances that can be used to train and test abuse detection systems in human-robot conversations. To this end, we designed an annotation scheme to label non-abusive and abusive human utterances, in particular, discriminating general offense from sexual abuse and hate speech. We've manually annotated $\sim$1000 utterances with agreement of 70%, and we've evaluated existing approaches that deal with abuse detection. We then chose the approach that would best suit our purposes, and we've trained it on our own data. Furthermore, we've integrated our model in the main Alana chat-bot. Moreover, we've formulated four different mitigation strategies with appropriate responses, and we evaluated our system with more than 50 different real users. Results show that users significantly preferred our system with abuse detection and mitigation over the one without these features.

As a natural extension of this work, we plan to extend the data annotation, to study different machine learning approaches, and improve our mitigation strategies. In particular, we also considered to create an ensemble model of the Rasa and (Davidson et al., 2017) systems. With an additional dataset we can also imagine to test in the future the classification system proposed by (Founta et al., 2018) and a word2vec approach, e.g., (Bojanowski et al., 2016). Finally, the mitigation strategies need some work on understanding which category of answers, among the four we designed, works best for the received utterance.

Acknowledgments

We want to thank Prof. Verena Rieser and Dr. Amanda Cercas-Curry for the work provided for this project.

References

R Bacal. 1998. Defusing hostile customers workbook. *Institute for Cooperative Communication, Manitoba, Canada*.

Tom Bocklisch, Joey Faulker, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Sheryl Brahnam. 2005. Strategies for handling customer abuse of ecas. *Abuse: The darker side of humancomputer interaction*, pages 62–67.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Amanda Cercas-Curry and Verena Rieser. 2018. Ethical evaluation of conversational systems: Sexual harassment in the #metooalexa corpus. In under submission.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

A De Angeli, S Brahnam, P Wallis, et al. 2005a. Abuse: The dark side of human-computer interaction. In *Interact'05*, pages 91–92. Laterza.

Antonella De Angeli, Rollo Carpenter, et al. 2005b. Stupid computer! abuse and social identities. In *Proc. INTERACT 2005 workshop Abuse: The darker side of Human-Computer Interaction*, pages 19–25.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02):11–17.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM.

Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2018. A unified deep learning architecture for abuse detection. *arXiv preprint arXiv:1802.00385*.

Agneta Gulz, Magnus Haake, Annika Silvervarg, Björn Sjödén, and George Veletsianos. 2011. Building a social conversational pedagogical agent: Design challenges and. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices: Techniques and Effective Practices*, page 128.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2017. Ethical challenges in data-driven dialogue systems. *arXiv preprint arXiv:1711.09050*.

L Rowell Huesmann, Leonard D Eron, Monroe M Lefkowitz, and Leopold O Walder. 1984. Stability of aggression over time and generations. *Developmental psychology*, 20(6):1120.

Alexa, are you turning my kid into a jerk? https://www.usatoday.com/story/tech/nation-now/2017/06/07/alexa-you-turning-my-kid-into-jerk/375949001/. Accessed: 2018-04-13.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondřej Dušek, Verena Rieser, and Oliver Lemon. 2017. An ensemble model with ranking for social dialogue. *arXiv preprint arXiv:1712.07558*.

Karl Pearson. 1900. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

John B Pryor and Nora J Whalen. 1997. A typology of sexual harassment: Characteristics of harassers and the social circumstances under which sexual harassment occurs.

Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press.

Christina Salmivalli and Eija Nieminen. 2002. Proactive and reactive aggression among school bullies, victims, and bully-victims. *Aggressive behavior*, 28(1):30–44.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

George Veletsianos, Cassandra Scharber, and Aaron Doering. 2008. When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers*, 20(3):292–301.

Jina S Yoon and Karen Kerber. 2003. Bullying: Elementary teachers' attitudes and intervention strategies. *Research in Education*, 69(1):27–35.