Productive Online Discourse for Emergency Response

Gaurav Nuti,[2] Louis Penafiel,[1] Janelle Ward,[3] Abel Salinas,[2] Fred Morstatter,[2] Nathan Schurr,[1] Deirdre Kelliher,[1] Laura Cassani,[1] Robert McCormack[1]

[1]Aptima, Inc., Woburn, MA, USA
[2] University of Southern California, Information Sciences Institute, Marina Del Rey, CA, USA
[3] BRIO Solutions, Buckeye, AZ, USA

Abstract

Despite the potential for antisocial and counter-productive social media behavior–particularly in the context of humanitarian assistance/disaster response (HA/DR)–there is a paucity of automated methods to address it. Current methods focus primarily on detecting hate speech and banning problematic content. We propose an alternative strategy of using automated counter speech to focus not just on moderating uncivil behavior, but also the promotion of civil discourse. In this paper, we propose a novel framework to employ pre-trained language models to alleviate the bottlenecks in adoption of such counter speech, namely a lack of understanding on the dynamics of counter speech and a scarcity of well curated datasets, which are compounded in HA/DR settings. We utilize GPT LMs to create a conversational testbed to simulate online conversations where various approaches for counter speech and other content moderation methods can be evaluated. Additionally, we leverage BERT-based models to detect hate speech and other network and syntactic features to suggest the optimal strategy to employ. We also present empirical results on the experiments we have conducted which provide a proof of concept for the framework.

Introduction

There is a growing awareness among national and international policy makers, civil organizations, and social media companies of the damage caused by antisocial behavior. Several countries have created laws regarding it,[1] and social media platforms have their own policies to combat hate speech on their platforms[2]. However, most tools to combat hate speech involve banning or removal of content which risks violation of free speech. Much less effort has been focused on shifting the paradigm to develop capabilities for fostering cooperative engagement in digital interactions through automated methods. Although some social media platforms like Reddit use manual moderators to mediate discourse, it is time- and resource-intensive, and lacks the ability to scale across the many diverse platforms where online discourse takes place. Automated counter speech shows much promise in dealing with the above problems. However, the lack of well curated datasets impedes such efforts.

The above-mentioned challenges become more severe in the case of emergency response. Each new emergency situation brings with it time-sensitive information which has different dynamics. Additionally, the response may need to be relayed in multiple languages, therefore a multilingual approach is needed. Moreover, the stakes in emergency situations become higher where uncivil behavior could interfere with the disaster relief support. To better understand the real world challenges, a use case was referenced throughout the research effort, including toxic social media usage within a humanitarian assistance/disaster response (HA/DR) event.

Recently, advances in language models have demonstrated the prospect of realistic generated injects that can guide interactions. Transformer-based generative language models, such as GPT-2 and GPT-3 (Radford et al., 2019). Brown et al. (2020) have significantly advanced conversational agents' ability to generate contextually relevant content. Transformers have shown above-human-level expertise on the GLUE benchmark based on their bilingual evaluation understudy (BLEU) scores (Wang et al., 2018), and believability in their generations, even for long-form results (Destine-DeFreece et al., 2019). Furthermore, these models are language agnostic, and have been trained to produce generations in a variety of languages, such as Spanish and Portuguese. We use the English and Spanish[3] variants of GPT2 to create a "conversational testbed," where realistic conversations are simulated. We then plan to detect when the conversation becomes toxic using a multilingual variant of Bert(mBert). If conversation becomes toxic we will then serve automated mediation injects which will be fed as prompts in the conversational testbed. The impact of the inject on the conversation can then be qualitatively and quantitatively measured. We provide empirical results on initial experiments that provide a proof of concept for our system.

Related Work

We organize the related work around two themes: existing work on hate speech detection and mediation approaches.

[1] https://www.wikiwand.com/en/Hate_speech#/Hate_speech_laws

[2] https://transparency.fb.com/policies/community-standards/hate-speech/

[3] https://huggingface.co/DeepESP/gpt2-spanish

English hate speech detection has been vastly explored by the research community. Recent work has focused on deep learning techniques for English hate speech detection. Agrawal and Awekar (2018) show deep learning approaches, such as CNN, LSTM, and Bidirectional LSTM with attention, perform better than traditional machine learning techniques. In the multilingual setting, Aluru et al. (2020) experimented with multilingual embeddings, and transformer-based multilingual architectures. Some researchers leverage repositories like Hatebase (Silva et al., 2016), a crowd-sourced repository of hateful and derogatory words across many languages. Additionally, online tools exist for detecting toxicity such as Google's Perspective API[4].

We found a variety of mediation strategies and definitions of uncivil or toxic online behavior in the literature. For instance, Mathew et al. (2019) based their research on eight types of counterspeech. Findings revealed that the effectiveness of counterspeech strategies varied across communities. In our current research venture, the types of counterspeech explored by Mathew et al. guided discussions regarding how to mediate hate speech, and other forms of uncivil online behavior, within the HA/DR domain. Just as the types of mediation strategies vary, so do definitions of uncivil, anti-social, or toxic discourse. For example, Coe, Kenski, and Rains (2014) defined five different forms of incivility. These definitions informed the development of a taxonomy of uncivil behaviors in the current research effort.

Domain Use Case

A real-world HA/DR use case provided a foundation for development of hate speech detection methods, mediation approaches, and testbed generations. The use case reflects tensions that were evident on social media between Colombians and Venezuelans in 2019. The economic crisis in Venezuela led an increasing number of migrants and refugees to flee deteriorating conditions, while Colombia struggled to support both citizens of their neighboring country and its own people in need of assistance.

The following discourse posted on a Colombian newspaper's Facebook page exemplifies the friction between Colombians and Venezuelans. Note: comments were translated from Spanish to English in Facebook.

- I hope you never have to go hungry and watch your relatives die for lack of food and medicine
- That happens every day here sir
- Hey daughter, may God forgive you because you don't know what you're saying. You have to spend misery anguish shed tears for your family children spend hunger see every day your life
- You should study a little the history of Colombia before speaking what apparently you ignore and in Colombia every day people die of hunger and no politician does anything is that justice? Please. think before you speak

Such discourse was counterproductive in that the focus was on arguing rather than finding common ground regarding humanitarian needs on both sides of the border. If tensions had continued to rise, frustrated Colombians may have disrupted delivery of aid destined for Venezuela. Thus, toxic interactions on social media may negatively impact emergency relief operations. By mediating such discourse, it may be possible to create a more permissive environment for sharing of information critical to serving those in need of assistance.

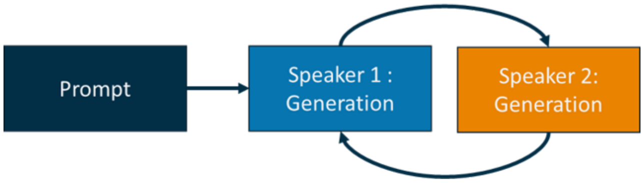

Figure 1: The workflow of the discourse generation for a two speaker generation.

Methodology

To achieve our goals of exploring and evaluating mediation methods, we have developed a discourse generation framework to simulate online conversations and track metrics at the utterance level. The workflow starts with topics containing background information. From these topics, prompts and initial utterances are created, to be used as prompts for unmediated discourse generation. Measures are then analyzed at three levels: participants, utterances, and conversations. Per our use case, hate speech detection algorithms are used to identify places for mediation. After choosing a mediation method, mediated prompts are created from the utterances of conversations prior to and including detected hate speech. With these prompts, mediated discourses, which contain injected mediations, are generated, then the generations and measures of discourse are compared for evaluation.

We focus on discourse generation, measures of discourse, and hate speech detection, saving discussions on other pieces–such as mediation strategies–for a later work. The discourse generation makes use of the recent advances in transformer-based generative language models, such that it is automated, adaptive, scalable, and multifarious. The goals of these generations are to create legible, germane, and believable conversation that have similar behaviors to those in the Domain Use Case discourse sample. The topics are changed by altering the prompts used to generate the conversations. Furthermore, we are investigating the injection of personas via prompts, as well as fine-tuning generation models, to add flavor and diversity to the generations. The workflow for the generations is shown in **Fig.** 1.

These conversations are then processed by the measures of discourse piece, to evaluate the effects of interjections in discourse tasks. The measures of discourse capture aspects, such as novelty and diversity of ideas; overall sentiment, tone, and emotion of conversations; and mediation interactions. We track these metrics on various strata of the discourse–specifically, utterance, participant, and conversation. Tracking these aspects of discourse allows us to assess the appropriateness of the mediation and to analyze its effects on the rest of the conversation or on the discourse as a whole. Because of the stochastic aspect of generative language models, we aim to test and evaluate the mediation by generating large amounts ($\mathcal{O}(100)$) of discourse with and

[4] https://perspectiveapi.com/

without the mediation. Then, we will aggregate the various metrics to come to a quantified evaluation of the effects of the mediator agent.

We use hate speech detection models to determine when to interject in the conversation. Due to the multilingual nature of conversations of interest, we are using Mbert, which is fine-tuned on Spanish and English hate speech datasets[5](Aluru et al., 2020).

Results

Here, we present a sample generation. For this generation, we use the base GPT2-XL generative language model. There are two speakers with names associated with them. Each conversation is composed of utterances and each utterance has a randomized length from one to three sentences.
In this first generation, the speakers are ‘Venezuelan’ and ‘Colombian’. The prompt is as follows:

> The President of Colombia participated in a press conference in response to the humanitarian aid that was halted at several Venezuelan border crossings. Citizens on both sides of the border posted comments on social media during and after the press conference. Some Colombians suggested that aid should be distributed within the country, such as for recent flood victims in Choco and Colombians suffering from famine in Guajira, instead of being delivered across the border to Venezuelan citizens. Citizens from both nations use social media platforms to debate the merits of the President’s claim.
> COLOMBIAN: Knock knock, sorry Mr. President, Mr. President, in the Choco they’re drowning. Didn’t anyone tell you that part of Choco is flooded and that they need help?
> VENEZUELAN: Why don’t you go and help dog?

This particular prompt provides three key aspects of the prompt that contributes context for the generations. First, relevant entities are included, such as the general locations (e.g., Colombia) and specific locations (e.g., Choco). Actors are also included, such as the speaker names (e.g., ’Colombian’) and political figures. Second, relevant topics are included, such as topics of conflict, specific events, and opinion attribution. Lastly, tone setting is used through scene setting and the addition of direct interactions between the speakers.
The subsequent generations from this prompt are:

> COLOMBIAN: No, no but I will
> VENEZUELAN: It’s okay, Mr. President, there are people who are already helping those people. What we want is food, medicine, water.
> COLOMBIAN: There are floods all over Venezuela.
> VENEZUELAN: But why don’t you ask them to distribute it where it belongs?
> COLOMBIAN: You know what, that’s not my problem.
> VENEZUELAN: You can do it.
> COLOMBIAN: Well, I’ll try to find something for myself. The President was asked about the incident with Colombian police officers detained by Venezuelan authorities while trying to cross into Colombia on July 26, 2016.

From these generations, we see various relevant entities, such as Mr. President and Venezuela. The conversation kept consistent positions and references to events discussed in the prompt. The tone setting established in the prompt assisted in steering the conversation to antagonistic questions.

While a thorough analysis is needed for validation, once a large number of generations are created, such analyses are untenable. We provide an example of a quantitative analysis that can be done using the measures of discourse. This analysis looks to compare some of the measures between 25 unmediated and 25 mediated discourse generations. For the unmediated discourse generations, we use the same prompt as earlier, but for the mediated generation, we add one more line at the end of the mediation, explicitly:

> MEDIATOR: This topic is getting divisive. We ask people to remember the rules to engage thoughtfully and to be respectful and considerate of other members.

For this analysis, we focus on the utterance level. The measures in this level include character counts, word counts, sentiment analysis, and emotion analysis. For the sake of simplicity, we focus on the sentiment scores. The sentiment scores are calculated using Vader (Hutto and Gilbert, 2014). The results shown in **Fig.** 2 are not meant to showcase that this mediation strategy is effective or not as this is preliminary. The values for these figures correspond to the average value for that line number over the 25 discourse generations. The sentiment scores are on a scale where negative values carry negative sentiment and positive values indicate positive sentiment. The sentiment scores correlating with toxicity is an initial hypothesis, which we have not yet validated. In the future, we will perform similar analyses and make use of conversation-level and participant level measures to test out different hypotheses.

Since most of the generations provided by GPT2-XL LM are not toxic or hateful, we experimented with using both prompts and fine-tuning the LM for hateful language. To compare we generated 1000 tweets using each method i.e using prompts, fine-tuning with an empty prompt, and fine-tuning with prompts. The prompts we used here are a list of three samples of Spanish hate speech that was found in social media. We translated the Spanish to English for the English prompt. For fine-tuning, we used the hateval dataset (Basile et al., 2019) to fine-tune the GPT2-M model for English and the GPT2-Spanish model for Spanish. We then used pre-trained mBert to detect the number of hateful tweets per 1000 generations, shown in **Fig.** 3. As we can see in both languages, we detect the most hate speech by combining fine-tuning with prompts. Fine-tuning the generation model biases it to generate more hate speech and become situationally aware. Additionally, we note that hate detection models are trained on tweets and might not work equally

[5] https://huggingface.co/Hate-speech-CNERG

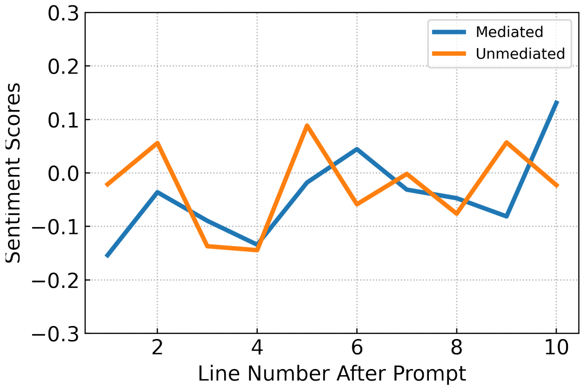

Figure 2: The workflow of the discourse generation for a two speaker generation.

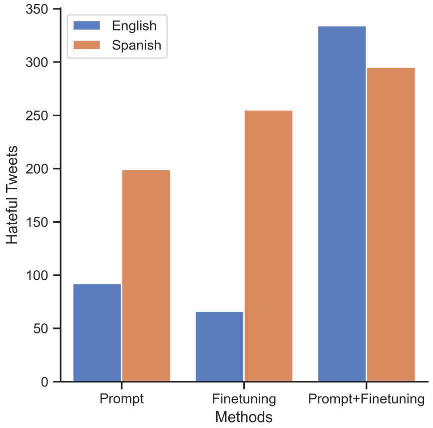

Figure 3: Comparing the number of generated tweets that were classified as 'Hateful.'

well on conversations, and might require us to fine-tune such models for conversations also. Furthermore, as shown before prompts provide the generation with more context, as we qualitatively noticed that when using prompts the hate speech is more topical. Consider the following examples:

- Venezuela is a nation that thrives on filth...
- #Venezuela has an immigrant problem that is far worse than immigrants coming from other countries

Conclusion

In this work we present an approach to counter toxic discourse in multilingual HA/DR environments. This approach works by identifying toxic discourse We demonstrate the capability of this system using our conversational testbed, finding that our method that leverages the prompt with finetuning yields the most hate speech.

Future work is to build this into a framework that can be deployed in response to emergent disasters. This will require culturally-sensitive mediation approaches. Furthermore, we will explore the space of meaningful metrics for identifying good conversations.

Acknowledgements

This project was sponsored by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR00112290021. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

Agrawal, S., and Awekar, A. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, 141–153. Springer.

Aluru, S. S.; Mathew, B.; Saha, P.; and Mukherjee, A. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Coe, K.; Kenski, K.; and Rains, S. A. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64(4):658–679.

Destine-DeFreece, A.; Handelsman, S.; Light Rake, T.; Merkel, A.; and Moses, G. 2019. Can gpt-2 replace a sex and the city writers' room?

Hutto, C., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media* 8(1):216–225.

Mathew, B.; Saha, P.; Tharad, H.; Rajgaria, S.; Singhania, P.; Maity, S. K.; Goyal, P.; and Mukherjee, A. 2019. Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media* 13(01):369–380.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners.

Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the targets of hate in online social media. In *ICWSM*.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.