VIEW POINT

A STEP TOWARDS CLEAN CONVERSATIONS

Abstract

Social Media being a medium for unrestricted communication and content, is not with its safety and regulatory concerns. Online abuse, privacy violation and bullying are at its peak and companies are doing their part to uphold digital civility and ensure clean conversations. This Point of View is a look at how companies and open source groups are ensuring cleaner conversations using AI and ML for safer and healthier online interactions.

Infosys®
Navigate your next

Digital Interactions & Civility

In the 20th Century, technology began to change rapidly. With the advancement of networks and internet as well as the proliferation of blogging and microblogging, the world witnessed the growth of social media platforms. Social Media platforms including networking sites, e-newspapers, blogs and the likes provided a venue for people to discuss common interests from films to religion and hobbies to politics. With discussions becoming a daily affair, melting the barriers of geographies and language, difference of opinion started to lead way to cyber bullying and abuse.

Many of these digital platforms are putting in community standards and measures to counter cyber bullying and abuse, yet many individuals refrain from sharing their thoughts and opinions fearing wrongful retribution. It is at this juncture, that we need to consider the effectiveness of the existing systems and think of how we can improve the same facilitating clean conversations.

Democratizing Internet without Digital Civic Engagement is Farce

Democratizing internet guarantees everyone has equal and unbiased access to the same content and opportunity. Internet represents a new transnational space for dialogue between people who would never otherwise encounter one another with great openness. This enables constructive and meaningful digital engagement that could bring our society and products to a better state. Social media platforms aim at making meaningful connections but that is possible only when interactions are authentic, genuine and clean.

Today, anyone who has participated in online conversations on social media knows the menace of online interactions. Digital gangsters are deterring many people from engaging online with vicious comments, hurtful, abusive, or derogatory language. As online conversations have mushroomed, so have their dangers rendering questions on democracy online.

The Pain of Conversing Online: Widespread Risk

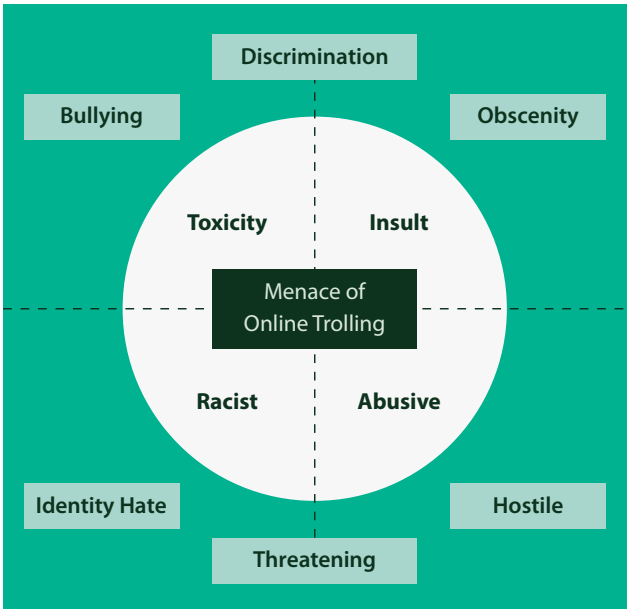
With the rise of social media platforms like WhatsApp, Chat Bots, Facebook, Twitter, Instagram, Slack etc. it has become our important transactional media to interact and exchange ideas.

Even when it is one of the game changers in communication and interaction media, it lacks high standard of Digital Civility. As per PEW Research, 27% of American internet users chose not to post something online after seeing someone being harassed.

The pain of online interactions and the toxicity around it has not limited social media platforms from being a favorite, exposing more and more people to cyber bullying and hate.

Therefore, there is an immediate need to curb negative online bullying and toxic comments to ensure the safety and mental health of social media users. Microsoft in one of its researches surveyed teens and adults across 22 countries, to examine the extent of negative behaviors and online interactions based on 21 internet risks. Based on the same, it developed Digital Civility Index (a measure of consumers' lifetime exposure to online risks such as Behavioral, Intrusive, Reputational and Sexual which are further categorized into several individual risks) and published a Global report on Type of Online Risks, Menace of Online Trolling and Digital Civility Index (DCI) across the world. The research revealed much about the year on year variations in toxic online conversations and called for sense of digital etiquette to act responsibly and connect online with dignity and respect.

Today's connected world demands a cultural shift to respect people's right to expression and diverse opinions. Enterprises, policy makers, companies and individuals need to come together to create awareness on digital civility, and should pledge to make the internet a better place for clean and meaningful discussions.



Industry Efforts Towards Clean Conversations

Today, many organisations have taken on the challenge to promote healthy and meaningful online conversations. Online social networking platforms like Facebook, Twitter, YouTube and LinkedIn have instated community guidelines and manual efforts to curb online toxicity. Alphabet's Conversation AI team, a research initiative founded by Jigsaw and Google (both part of Alphabet) are working on tools to help improve online conversation. So far, they have built a range of publicly available models served through the Perspective API, including one to ascertain toxicity in comments or content. Google has also come up with a set of "**AI ethics and principles**" to promote clean and healthy conversations.

Open source communities are also catching up with the clean conversation wave. Even though there are no open source softwares or APIs that are currently available, communities are helping each other by providing datasets for training AI/ML models. To help improve the accuracy of the models, many big tech companies and social media sites have started releasing huge datasets for training machine-learning models that detect toxic behavior. Kaggle has many such datasets like insincere questions by Quora and Toxic Wikipedia comments by Jigsaw. In addition to the Kaggle dataset, one can also make use of Reddit comments data, Fake News data collection and many more

Initiatives by Infosys

Resonating with the social need for cleaner conversations and a safe digital space, Infosys is working towards building safer and responsible chatbot conversations. Most of present day algorithms are capable just enough to perform binary classification of a conversation with the chatbot, i.e to identify whether the conversation is happy, sad or neutral and then provide a summary of the same to the agent to whom it would be transferred. To bring in more intelligence to this task and help weed out objectionable content and raise alarms, NIA Chatbot Platform is working on identifying specific sentiments. With this initiative, enterprises can ensure cleaner chatbot conversations and better agent summaries. The service will be made available as both a part of NIA Chatbot Platform and a micro service.

Recently, the Conversational UI research team of the NIA Chatbot Platform participated in Kaggle Toxic Comment Challenge to create a high accuracy model for detecting toxicity levels, which can then be utilized in production.

Here is a glimpse of how the team detected toxicity levels.

Examples of Sample sentences and its Toxicity levels

You should Kill that stupid person	
Toxic	0.969
Severe Toxic	0.136
Obscene	0.528
Threat	0.557
Insult	0.621
Identity Rate	0.065

This is example of clean comment	
Toxic	0.006
Severe Toxic	0
Obscene	0.001
Threat	0
Insult	0.001
Identity Rate	0

Analysing the sentence, "You should kill that stupid person" it is labeled Toxic with a confidence score of 0.96, while the model tags the sentence "This is example of clean comment" toxic with a very low confidence score of 0.006.

As per the devised solution, one can easily include a mechanism in the comment section of social media platforms or in Conversational platforms to keep track of toxicity, hate, threat, obscenity, or insult. This will help score user interactions following which the comments or users can be monitored with some Auto-generated warning or direct deletion.

Here is a step by step explanation of the Data Pipeline followed:

Data Preprocessing

- Data cleaning (removing html tags, ip address etc.)
- Data preprocessing (expanding apostrophes, lemmatization etc.)
- Converting to sequence data for model inputs

Pretrained Emeddings

- Fasttext and Glove pretrained embeddings, which are trained on large text corpus

Deep Learning Models

- RNN and attention based models
- Different model architectures and hyperparmater tuning
- 10 fold cross validation for better accuracy

Submission and Result

- Final submission achieved 0.9860 ROC AUC score (First rank ROC AUC score 0.9885)
- Ranked in Top 24 % (Among 4551 teams)

Future Improvements

- Utilizing emoji data present in comments
- Experiment with new state of state models like bert
- Error analysis for understanding model

Improving Online Conversations with Technology

Machine Learning has been used in the past for detecting many kinds of online frauds and abuses. Given the large amount of training data available, we can train ML models to differentiate between normal and toxic behavior.

PhotoDNA from Microsoft is one such example, which tries to find a match from previously identified illegal images. PhotoDNA is being actively used for detecting images of child exploitation. It helps in removing pornographic images, abusive text data and propaganda videos.

One can also leverage Google Vision API to help in the same. Here is an example of how Google vision API can be utilized to remove toxic content.

Vision API response json contains Safe Search column, which can identify adult, spoof, medical and violence in images. Enabling the use of Safe Search to flag images that may contain adult and violent content.

Vision API also identifies text content present in images, which can be utilized further to predict toxicity using text based models.

Fig 1.1: Screenshot of a picture being assessed by the application

Machine learning models are not hundred percent accurate, but It can be utilized to clean online toxic content and abusive users at scale. Considering a news site that generates 1000 user comments per second, to maintain healthy conversations the site will need to hire huge number of people to regulate abusive comments that surface every second. In such instances, machine-learning models can be used to flag such abusive or toxic comments and trigger warning messages to such users; it can be further used to invoke actions against violators such as Review, Delete, Hide or Ban.

By improving existing Machine Learning models and incorporating a bit of Human supervision, organizations will be able to improve the standard of online conversation and help ensure safer and civil social platforms.

Conclusion

Regulated Digital conversations, ensuring high standard of digital civility, standards guiding online interactions along with AI powered techniques can help in making social media platforms safer and inclusive. To tap the maximum potential of a democratically sound digitally connected world, there requires an overall cultural shift to respect diverse opinions and differences. It is also equally important to educate internet users not only about their rights to safeguard themselves but also about their duty to be sensitive to others opinions and differences to ensure healthy online interactions. The time has come where every big and small enterprises and individuals need to step up to contribute towards Civil Digital Interactions in order to maintain the relevance and contributions of the digital space.

About the Author

Amit Kumar is an Architect working with the Conversation UI Research Team. Amit specializes in Artificial Intelligence, NLP, Machine Learning, Deep Learning, Cognitive, Mobile, Social, Wearable and converged Digital Experiences. Amit leads a team in the Innovation & product R&D space with passion for continually creating new business value in entrepreneurial setting. Amit is involved in conceptualizing and generating intellectual property, and building product lines in the areas of Artificial Intelligence.

Co-Author

Dharmendra Choudhary is a Specialist Programmer working with the Conversation UI Research Team at Infosys Center for Emerging Technology Solutions. He specializes in Deep Learning, NLP, HSI(Human System Interaction) and AI models in production.

Reference:

- https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
- https://medium.com/the-false-positive/the-challenge-of-identifying-subtle-forms-of-toxicity-online-465505b6c4c9
- https://perspectiveapi.com/#/
- https://www.microsoft.com/en-us/digital-skills/digital-civility?activetab=dci_reports%3aprimaryr6
- https://www.blog.google/technology/ai/ai-principles/
- https://www.facebook.com/notes/facebook-security/improvements-in-protecting-the-integrity-of-activity-on-facebook/10154323366590766/
- https://www.theguardian.com/technology/2019/feb/18/facebook-fake-news-investigation-report-regulation-privacy-law-dcms
- https://cloud.google.com/blog/products/gcp/filtering-inappropriate-content-with-the-cloud-vision-api

For more information, contact askus@infosys.com

Infosys®
Navigate your next