Trust & Safety

Engaged safety measures to curb hate speech and enhance end-user experience for an internet company

Situation

- Genpact manages hate speech on the client platform for NA
- 53% of Americans are a victim of hate speech & 37% of Americans are a part of severe hate crimes
- 31% yearly increase in hate crimes

Challenges

- AI can identify very limited content associated on hate speech due to its subjective and ambiguous nature of content. Thus, human intervention is absolutely necessary
- Battling human bias while moderating hate speech content
- PR issues arising due to False Positives – taking content off the platform when it is not hate speech or posted as a part of a discussion
- Changes in the community perspective where aspects considered an attack is no longer an attack or vice versa

Solutions

- Change in operating model to promote specialization on violation queue type
- Framework designed to create an ecosystem of support for teams to facilitate calibrations, clarifications and training interventions
- Digital dashboard created to enable real-time coaching and overcome delays due to dispute and overturn
- Wellness centered performance measurement for teams to balance wellbeing with accuracy and productivity drivers
- Insights generated towards simplifying policy and protocol updates as a measure to mitigating impact

Impact generated

- 4% increase in accuracy
- 6% decrease in False positives and False negatives
- Client moved 90% of volumes currently generated on the platform to Genpact

Curtailed cyber bullying and harassment through content moderation for a community platform

Challenges

- Impact on community in the form of Online Bullying and Harassment leading to life threatening events
- Market and cultural nuances in identifying bullying and harassment attacks. E.g., intent
- Identifying the target
- Determining age of victim

Solutions

- As subject matter experts, we identify the following
  - Target
  - Attack
  - Purpose of the post (awareness or hate/bullying)
- Creating market specific term lists and cheat sheets
- Methodical guidelines to identify age of poster and victim

Impact generated

- Creation of market specific cheat sheets, slur lists and term lists created to benefit all vendors
- Over 50 insights provided to make the policy more relevant and reduce ambiguity

Moderated sensitive content for a major technology company

Situation

- There is a global need to ensure safe environment to the users by actively reviewing suicidal and self-harm related content on the platform and provide necessary help to the end user

Challenges

- Educate content reviewers to comprehend ambiguous posts of user statement/comment/situation/intent of the user mentioning about all self-harm
- Quick decision-making ability and high sense of ownership as this is very critical to a user’s life

Solutions

- Rigorous training / refresher sessions imparted to be alert and perceptive of human behavior that prompts and confirms towards a possible attempt by users on suicide and self-harm
- Escalations to the client legal enforcement team who further work as per the law-and-order policies for the users from each company
- The TAT for all such critical queues is 1hr to escalate they find critical and escalation worthy
- Regular insights provided to the client to improve the policy guideline
- QA/Trainer/SME aligned to each reps to ensure that they are coached to take quick decisions

Impact generated

- Quick turnaround time in escalating to client team within 30 mins.
- 37 insights provided to improve the policy guideline in 2019
- In 2018, of the 8062 cases escalated, 3653 were lifesaving cases
- In 2019, of the 10406 cases escalated, 3502 were lifesaving cases