Jigsaw's *Perspective* on Combating Internet Harassment
By: Rapid Access International, Inc.
February 2017

始めに

インターネットは、想像しうる事項のほぼすべてに関して、アイディアの共有や意見の発言を行う最大の媒体となった。また、インターネットには、嫌がらせ（ハラスメント）や感情的な反応の挑発を意図した、根拠がなく悪意を持ったコメントを掲載するような人でも溢れている。このような人は「インターネット・トロール」とも呼ばれている。Data and Society Research Institute（データ・社会研究所）[1]によると、米国人の半数がインターネット・ハラスメントの被害経験がある。また、このような問題への啓発を目的として、「Hack Harassment」のようなハラスメント対策団体が複数設立されている。[2] ハラスメント対策団体の主唱者が技術系企業に対し、利用者を保護する製品の開発を要望したことに応え、Google 系列の技術インキュベータである Jigsaw と、Google の乱用対策技術チーム（Counter Abuse Technology Team）が新しいソフトウェア「Perspective」の開発を開始した。

「Perspective」ツール

Alphabet（Google の親会社）の子会社である Jigsaw は 2 月 23 日、インターネット・ハラスメント対策を目的に「Perspective」を立ち上げた。ここで用いられるアプリケーションプログラミングインターフェース（API）は、Jigsaw の「Conversation AI」（会話人工知能）というインターネットスイートのツールであり、ソーシャルメディアやウェブフォーラムでのハラスメント行為を捜し当てるのに用いられる。[3]「Perspective」は適応性の人工知能（adaptive artificial intelligence）を用いて、インターネット・コンテンツをリアルタイムで精査し、悪意のある発言を検索したうえで、独自の「有害度」に基づきフラグ付け、等級付けがされる。利用者は、ウェブフォーラムの「有害度」[4]を参照でき、それに参加したいかを自ら決定することができる。

---

[1] データを中心とする技術開発から生じる社会的・文化的問題に関して研究を行う、独立・非営利研究機関。本部はニューヨーク市に所在。https://datasociety.net/about/

[2] https://www.washingtonpost.com/news/the-switch/wp/2017/02/23/google-fights-online-trolls-with-new-tool/?utm_term=.4b44401469c0

[3] http://wccftech.com/google-justic-league-launch-war-on-trolls/

[4] 本稿での「有害」とは、「個人を議論から外させる理由となりうるような、乱暴、無礼、不合理なコメント」を指す。**Toxic** in this article's context is defined as: "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." http://mashable.com/2017/02/23/google-jigsaw-moderation-tool/#036CQcgtbkqM

ウェブアドミニストレータは「有害度」統計を使って、フォーラムからの個人の締め出しや、悪意のあるコメントや投稿の削除ができる。

開発

Jigsaw の開発チームは、ニューヨークタイムズ、ウィキペディア、ハラスメント対策団体に投稿されたコメント数百万件をマイニングし、ハラスメントの事案を検索した。Google 従業員数千人がこのようなコメント欄を綿密に調査し、投稿が「有害」か「非有害」かを記録していった。Jigsaw のチームはこの結果を用いて「有害度」等級を開発し、これが「Perspective」プロジェクトの中核となっている。[5]

「Perspective」は、適応性の人工知能を用いて構築されており、使用する度に「学習」していくシステムとなっている。あるコメントが「有害」と判定されると、データが収集されスキャニングアルゴリズムが高度化されていく。「Perspective」が誤って「有害」の投稿と判定した場合には、利用者がその旨報告できるようになっており、学習システムが今後に向けて再訓練されていく。Jigsaw の発表によると、現在「Perspective」の確度は 92%、誤謬率は 10%である。[6]「Perspective」は今後時間をかけて、ハラスメント認識の効率、精度を向上させていくことになる。

検閲の懸念

「Perspective」はこのような「有害」な投稿を排除するのではない、という点には注意が肝要である。Jigsaw の Jared Cohen 社長は、人間の判断を回避したくはないとの考えであり、目標は、判断が容易にできるような印付けをして、モデレータが「有害」な投稿の取り扱い方を決定できるようにすることである、としている。ウェブアドミニストレータは悪意のある投稿はどれかという「Perspective」の判定を見ることができ、コンテンツを除去するかの決定はこのようなウェブアドミニストレータに任されている。

どのコンテンツを削除すべきかの決定には、問題に陥る危険も伴う。一部のコメントを選んで削除する行為は、言論の自由という基本的原則を蹂躙する検閲となりうるためだ。オンライン上の市民の自由及び人権の擁護を目的とした非営利法人である「Center for Democracy & Technology (CDT)」において「Free Expression Project（表現の自由プロジェクト）」を指揮する Emma Llanso 氏は、このツールの使用に関して警鐘を鳴らしており、「自動検出システムを導入すれば、時間・資源を費やしてまで誤認を確認しないままに全件削除するという選択肢の可

[5] https://www.wired.com/2017/02/googles-troll-fighting-ai-now-belongs-world/
[6] http://wccftech.com/google-justic-league-launch-war-on-trolls/

能性を開くことになりかねない」と主張している。「Perspective」の代表者は、現在のウェブ・モデレーティング・システムは、コンテンツのブロックまたは除去のみを目的としたものであり、「Perspective」は会話の質を向上させるためのツールと捉えるべきだ、と述べている。

終わりに

「Perspective」は、Jigsaw の「Conversation AI」（会話人工知能）スイートのツールの一つに過ぎず、現状では完成には程遠い状態だ。複雑な情報環境からハラスメントが発生する新たな経路が形成していくにつれ、「Perspective」が学習し会話環境に順応できる能力は重要性を増していく。「インターネット・トロール」と戦うことで、個人がハラスメントを恐れずに会話に参加することが促され、議論やアイディア創出が活発に行われるような、開かれた建設的な環境の構築が望まれる。

Jigsaw's *Perspective* on Combating Internet Harassment

By: Rapid Access International, Inc.

February 2017

Introduction

The internet has become the largest outlet for sharing ideas and voicing opinions on nearly any subject imaginable. The internet is also full of people who post unwarranted malicious comments that are aimed to harass and provoke emotional responses – also known as "internet trolls." The Data and Society Research Institute have found that 50% of Americans have been a victim of internet harassment, and anti-harassment groups such as Hack Harassment have been founded to bring awareness to this topic.[7] In response to leaders from anti-harassment groups requesting technology companies to develop a product to protect people, Jigsaw and Google's Counter Abuse Technology Team began developing new software called *Perspective.*

Perspective Tool

On February 23rd, Jigsaw - subsidiary of Alphabet (Google's parent company) – launched *Perspective* to combat internet harassment. This API is a component of Jigsaw's 'Conversation AI' which is an internet developer suite of tools that are to be used to locate harassment on social media and web forums.[8] Using adaptive artificial intelligence, *Perspective* scans through internet content in real-time, and searches for malicious remarks which are then flagged and graded based on it's 'toxicity' level. Users are able to see how 'toxic[9]' a web forum is, and decide for themselves if they would like to participate. Web administrators can use the 'toxicity' statistics to remove individuals from forums, and/or delete malicious comments or posts.

Development

The Jigsaw development team mined through millions of comments that were posted on the *New York Times,* Wikipedia, and from anti-harassment advocacy sites searching for instances of harassments. Thousands of Google employees sifted through these comment sections and marked posts as either 'toxic' or 'not toxic.' Using their findings, the Jigsaw team developed a 'toxicity' scale which is the backbone of the *Perspective* project.[10]

Perspective is built with adaptive artificial intelligence which allows the system to "learn" the more it is used. When a comment is flagged as 'toxic,' the data is collected to help refine the scanning algorithm. Users are also able to report instances where *Perspective* has incorrectly flagged a 'toxic' post, and the learning system will be retrained for future occurrences. Currently, Jigsaw claims that *Perspective* has

---

[7] https://www.washingtonpost.com/news/the-switch/wp/2017/02/23/google-fights-online-trolls-with-new-tool/?utm_term=.4b44401469c0

[8] http://wccftech.com/google-justic-league-launch-war-on-trolls/

[9] **Toxic** in this article's context is defined as: "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." http://mashable.com/2017/02/23/google-jigsaw-moderation-tool/#036CQcgtbkqM

[10] https://www.wired.com/2017/02/googles-troll-fighting-ai-now-belongs-world/

an accuracy rate of 92% and a 10% false positive rate.[11] Over time, *Perspective* will become increasingly more efficient and accurate at recognizing instances of harassment.

Dangerous to Censor?

It is important to note that *Perspective* does not remove these 'toxic' posts. The President of Jigsaw, Jared Cohen, said that he doesn't wish to bypass human judgement - the goal is to "flag low-hanging fruit" for the moderators to decide how to handle 'toxic' posts.[12] The decision to remove content is bestowed to the web administrator who can see which posts were identified as malicious by *Perspective*.

There is a danger that it can become a slippery-slope when deciding which content should be deleted – removing select comments is censorship that disrupts the fundamental principles of freedom of speech. Emma Llanso of the Free Expression Project has shown caution in using the tool, stating that "an automated detection system can open the door to the delete-it-all option, rather than spending the time and resources to identify false positives."[13] *Perspective* representatives note that current web moderating systems are designed to only block or remove content; *Perspective* should be viewed as a tool to advance the quality of conversation.

Conclusion

Perspective is only one tool from Jigsaw's Conversation AI suite, and is still far from being perfected. *Perspective's* ability to learn and adapt to the conversation environment will be important as the complexities of the information environment create new avenues for harassment to occur. The hope is that fighting 'internet trolls' will promote individuals to join in on conversations without fear of harassment, thereby creating a more open and constructive environment for topical debates and ideas to flourish.

Further Exploration of *Perspective*: https://www.perspectiveapi.com/

[11] http://wccftech.com/google-justic-league-launch-war-on-trolls/
[12] https://www.washingtonpost.com/news/the-switch/wp/2017/02/23/google-fights-online-trolls-with-new-tool/?utm_term=.4b44401469c0
[13] https://www.wired.com/2017/02/googles-troll-fighting-ai-now-belongs-world/