2018 年 4 月 13 日

報道関係者様各位

反レイシズム情報センター（ARIC）
代表 梁英聖

Twitter 空間におけるヘイトスピーチに関して、Twitter 社の募集する健全性指標測定への提案書に応募しました！

拝啓

時下ますますご清栄のこととお喜び申し上げます。平素はなにかとご厚情にあずかり誠にありがとうございます。

さて、このたび反レイシズム情報センターでは、Twitter 社が募集する健全性指標測定への提案書に応募いたしました。[i]

昨今 Twitter 空間におけるヘイトスピーチが問題となっています。世界的には Twitter や Facebook などの SNS において差別による攻撃的な投稿や、それを繰り返すアカウントの凍結など規制が強化されています。

しかし、日本においてはそのような対策は進まず、深刻なヘイトスピーチが放置されていることが度々問題となってきました。そこで、反レイシズム情報センターでは、この度 Twitter 社が公募する健全を測定する方法に応募いたしました。これは**インターネット上の差別的投稿を効果的に規制する際の一つの方法を提案**しています。

反レイシズム情報センターでは人種差別撤廃条約に違反する、公人の差別発言を記録した「政治家レイシズムデータベース」を作成しています。[ii]これらのアーカイブを無償提供し、当団体と他の国際人権法やマイノリティ問題の専門家、そして IT 専門家が協同して差別的投稿を検索するプログラムを開発することにより、**Twitter 社の定めるルールに違反する投稿を自動的にリストアップすることが可能になります。本案が日本におけるインターネット空間のヘイトスピーチ規制の議論の一助となることを期待しております。**

つきましては、本提案についてなにとぞ貴媒体でも取り上げていただきたく、ここにご案内する次第です。概要は下記の通りです。

Twitter 社の健全性指標測定への提案（※ツイッター社の応募フォームに沿ったものとなっています）

■Propose and define a health metric that Twitter could use to measure itself. *（ツイッター自身を測定するのに使える健全性指標を定義・提案してください）

【健全性指標1】ルールに反する差別の撤廃度：やむなく発生する差別・差別煽動のうち、ルールに違反するものを、ツイッター社がどれぐらい削除・アカウント凍結できているどうか。

※差別撤廃は世界人権宣言や人種差別撤廃条約・女子差別撤廃条約など各種国際人権条約が義務付けているものである。

■How would Twitter capture, measure & evaluate this health metric? *（どのようにツイッター社はその健全性指標を取得し、測定し、発展させることができますか？）

【指標の取得法】

〇現状

現状ではツイッターにはごく形式的な差別禁止ルールはあるものの、悪質な差別ツイートとそれを繰り返すアカウントが大量に放置されている。しかも差別禁止ルールに基づいてどのように差別か否かを判断しているか、またツイートを削除しているかが非公開であるため、ルールに違反する差別の撤廃度はまったく測れない。

〇指標の導入

ルールに違反する差別の撤廃度を健全性指標としてツイッターが取得するには、

A）実際に放置されている大量の差別を確実に撤廃すること、

B）その差別撤廃の方法と結果についてできるかぎり公開すること、

の 2 つが必要となる。

以下、その具体的方法を提案する。

1．過去のヘイトスピーチを大量に集めたアーカイブから、ルール違反のツイートを検索することを可能にする重要データを検出し、それを元に自動的にルール違反ツイートが疑われるツイートを検索するシステムを開発する。その実施主体として ARIC だけでなく、他の IT や情報科学の専門家・研究者、そして他の国際人権法・多文化主義・マイノリティ問題に詳しい専門家・NGO・研究者と協同するプロジェクトチームをつくる。具体的には、

1）私たち反レイシズム情報センター（ARIC）が運営する政治家レイシズムデータベース（https://antiracism-info.com/database_home/　）には人種差別撤廃条約第 1 条でいう「人種差別」に反する公人（政治家を中心に）の差別言動が 4358 件（2018 年 4 月 7 日現在）登録されている。ARIC はこれらヘイトスピーチのアーカイブを無償で提供する。

2）上記データを AI に読み込まるなど適切な方法により、ルールに違反すると強く疑われるツイートを自動的に検索するのに活用可能な次の 3 つの指標、①差別語・②複合検索すれば違反ツイートを検索可能な差別キーワード・③典型的差別パターンを検出する。

3）上記①②③を活用して自動的にルール違反が強く疑われるツイートを検索するプログラムを開発する。

※ARIC だけでなく、他の IT や情報科学の専門家・研究者、そして他の国際人権法・多文化主義・マイノリティ問題に詳しい専門家・NGO・研究者と協同するプロジェクトチームを発足させ実施主体とする。

2．上記プログラムを用いすべてのツイートから自動的に定時間ごとに（1 時間毎が望ましい）ルール違反が強く疑われるツイートをリストアップする「イエローリスト」を作成する。

3．「イエローリスト」のうち、プロジェクトチーム内の国際人権法や差別問題の専門家・NGO（あるいはその研修を受けたスタッフ）が、ルール違反とは言えないものを除外し、残りのルール違反のツイートをリスアップした「レッドリスト」を作成する。

4．ツイッター社は「レッドリスト」を参考にしてルール違反のものを自社の基準で削除する。

【指標の測定法】

1. 上記の方法じたいを公表することは、従来不透明であったツイッター社の差別撤廃に取り組む客観的な方法を可視化させツイッター社の差別撤廃への意思を公にするという意味で、差別の撤廃度という指標を測定する基本条件となるだろう。

2. 差別の撤廃度という指標を量的に測定することは簡単ではないが、次のことが考えられる。

1）毎週（毎日あるいは毎月）ごとに「イエローリスト」「レッドリスト」そしてツイッター社が実際に削除・凍結したアカウント数の統計データを公表する。

2）「「レッドリスト」÷「イエローリスト」」あるいは「ツイッター社の削除数÷「レッドリスト」」などの数値も公表する。これは差別の撤廃度という指標を測る参考値になるだろう。

【指標の発展法】

1. 「イエローリスト」「レッドリスト」「「レッドリスト」÷「イエローリスト」」「ツイッター社の削除数÷「レッドリスト」」の各数値はそれじたいの高低が直ちに差別の撤廃度を表すのではない。

2. しかし各データを短評する付加情報とともにそれらデータは差別の撤廃度の指標を測定する重要な根拠となるだろう。

3. 各データの短評をたとえば ARIC や他の人権 NGO や研究者に月報として公表するという案がありうる。

4. 各データが公表されれば、ツイッター上やその他でそれらデータの意味・文脈についてツイッターの差別の撤廃度に関する議論が起こるだろう。その議論じたいが一つの差別の撤廃度を測る指標になるのではないか。

（以下、第二・第三・第四の指標の提案について同じ問い）

■Propose and define a health metric that Twitter could use to measure itself. (metric #2)

【健全性指標2】ルールに反する極右の差別煽動活動の撤廃度：やむなく入り込んでくる極右の差別煽動活動のうち、ルールに違反するものを、ツイッター社がどれぐらい削除・アカウント凍結できているどうか。

■How would Twitter capture, measure & evaluate this health metric? (metric #2)

健全性指標1を導入することを前提としている。

そのうえでさらに、悪質な差別を繰り返すアカウント・極右グループの実態をよく知る NGO や専門家の意見を参考にして、極右アカウント・活動を撤廃する。これは前述プロジェクトチーム内の国際人権法・マイノリティ問題の専門家・研究者と協同で極右活動のヘイトウォッチ hate watch チームを発足させ、そのチームが収集した極右活動のデータから、

1）社会的影響力のある極右アカウントをリストアップする「極右リスト」を作成し、

2）そのうちツイッターのルールに違反するツイートあるいは活動に利用しているアカウントを選び出した「NG 極右リスト」を作成する。

3）「NG 極右リスト」はツイッター社に提供され、ツイッター社がアカウント凍結を行う参考にする。

4）本指標の測定・公開は「極右リスト」「NG 極右リスト」およびその中からツイッター社が凍結したアカウント数などを定期的に公表することで達せられるだろう。

■Propose and define a health metric that Twitter could use to measure itself. (metric #3)

【健全性指標3】。違反となる差別のガイドラインを NGO や専門家・研究者と協力して作成し、そのガイドラインに違反しているツイッターがどれぐらいあるかを公表する。

■How would Twitter capture, measure & evaluate this health metric? (metric #3)

FIFA がロシアワールドカップの差別監視プログラムで提携している国際 NGO である FARE は独自の差別ガイドラインを作成し公表している。そのガイドラインに反している差別が発見された場合、24 時間以内に FIFA に通報し、FIFA が対処することになっている。

これにヒントを得て、違反となる差別のガイドラインを NGO や専門家・研究者と協力して作成し、そのガイドラインに違反しているツイッターがどれぐらいあるかを公表する。

指標1と2と組み合せて運用すればよい。

■Link to relevant papers you've published (link #1) *（あなたが公表した関連する論文のリンク先）

梁英聖, 2016,『日本型ヘイトスピーチとは何か――社会を破壊するレイシズムの登場』, 影書房.

■Link to relevant publications (link #2)（関連する論文のリンク先）

高史明, 2015,『レイシズムを解剖する――在日コリアンへの偏見とインターネット』, 勁草書房.

以上

提案代表者プロフィール
梁英聖 : 一橋大学大学院言語社会研究科博士後期課程。反レイシズム情報センター（ARIC）代表。2016 年 12 月影書房より『日本型ヘイトスピーチとは何か――社会を破壊するレイシズムの登場』を出版。在日 3 世として日本社会の反レイシズム運動に取り組み、レイシズムの研究活動にも従事している。

なお、この件についてのお問い合わせは、下記までお願いいたします。

敬具

記

代表：梁英聖
E-mail：contact@antiracism-info.com

以上

[i] Twitter 社の募集要項は以下の URL の通りです。
https://blog.twitter.com/official/ja_jp/topics/company/2018/0302health.html
応募フォームは以下の URL の通りです。
https://docs.google.com/forms/d/e/1FAIpQLSdYhNZnbkmNlAsw0kFXgRFD9Z6XAO4rJEazTS_skGdN2YYpmA/viewform

[ii] 反レイシズム情報センター「政治家レイシズムデータベース」をご参照ください。
https://antiracism-info.com/database_home/