資料1－1

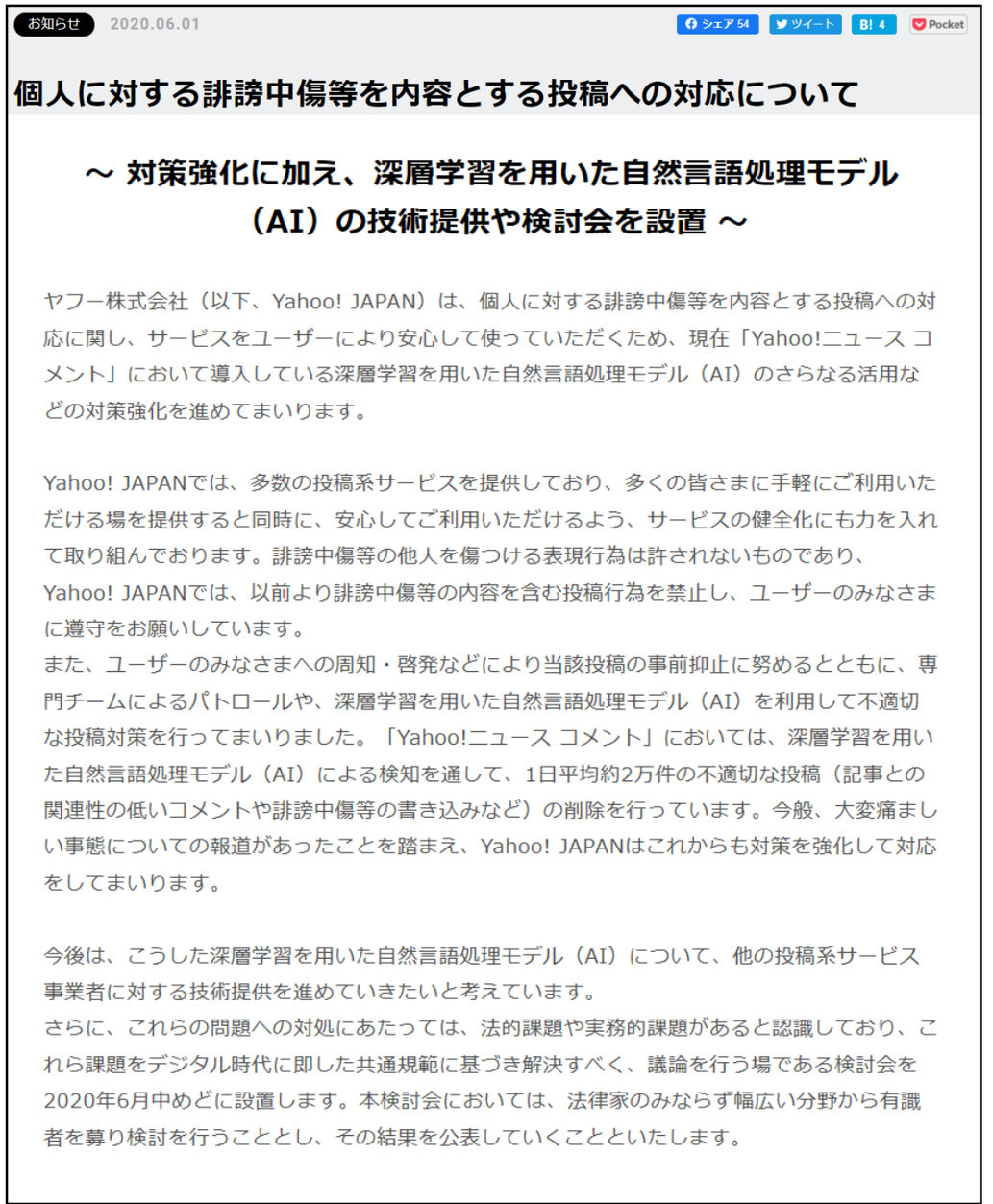
YAHOO! JAPAN

総務省プラットフォーム研究会

Yahoo!の取り組みについて

2021年2月25日
ヤフー株式会社

誹謗中傷投稿への対応



お知らせ　2020.06.01　シェア 54　ツイート　B! 4　Pocket

個人に対する誹謗中傷等を内容とする投稿への対応について

～ 対策強化に加え、深層学習を用いた自然言語処理モデル（AI）の技術提供や検討会を設置 ～

ヤフー株式会社（以下、Yahoo! JAPAN）は、個人に対する誹謗中傷等を内容とする投稿への対応に関し、サービスをユーザーにより安心して使っていただくため、現在「Yahoo!ニュース コメント」において導入している深層学習を用いた自然言語処理モデル（AI）のさらなる活用などの対策強化を進めてまいります。

Yahoo! JAPANでは、多数の投稿系サービスを提供しており、多くの皆さまに手軽にご利用いただける場を提供すると同時に、安心してご利用いただけるよう、サービスの健全化にも力を入れて取り組んでおります。誹謗中傷等の他人を傷つける表現行為は許されないものであり、Yahoo! JAPANでは、以前より誹謗中傷等の内容を含む投稿行為を禁止し、ユーザーのみなさまに遵守をお願いしています。
また、ユーザーのみなさまへの周知・啓発などにより当該投稿の事前抑止に努めるとともに、専門チームによるパトロールや、深層学習を用いた自然言語処理モデル（AI）を利用して不適切な投稿対策を行ってまいりました。「Yahoo!ニュース コメント」においては、深層学習を用いた自然言語処理モデル（AI）による検知を通して、1日平均約2万件の不適切な投稿（記事との関連性の低いコメントや誹謗中傷等の書き込みなど）の削除を行っています。今般、大変痛ましい事態についての報道があったことを踏まえ、Yahoo! JAPANはこれからも対策を強化して対応をしてまいります。

今後は、こうした深層学習を用いた自然言語処理モデル（AI）について、他の投稿系サービス事業者に対する技術提供を進めていきたいと考えています。
さらに、これらの問題への対処にあたっては、法的課題や実務的課題があると認識しており、これら課題をデジタル時代に即した共通規範に基づき解決すべく、議論を行う場である検討会を2020年6月中めどに設置します。本検討会においては、法律家のみならず幅広い分野から有識者を募り検討を行うこととし、その結果を公表していくことといたします。

■2020年春、誹謗中傷投稿問題の議論が過熱

■同年6月 1日　ヤフープレスリリース掲載

「プラットフォームサービスの運営の在り方検討会」を設置
個人に対する誹謗中傷等を内容とする投稿への対応について、外部有識者の意見も広く取り入れながら、ヤフーの対応方針について協議するもの。

■2020年12月23日　プレスリリース
上記検討会における議論を受けた提言書を受けて、ヤフーの対応方針を決定。

■2021年春以降～　順次対応をすすめる。

「プラットフォームサービスの運営の在り方検討会」概要

目的

個人に対する誹謗中傷等を内容とする投稿への対応を検討する

検討事項案

（1）目指すCGMサービスの方向性について

（2）AIを用いた対策の効果と妥当性について

（3）措置基準や実績に関する透明性の確保について

（4）業界横断の対応について

「プラットフォームサービスの運営の在り方検討会」概要

構成員（敬称略）

■ 座長

山本 龍彦（慶應義塾大学大学院法務研究科 教授）

■ 委員

小川 一（毎日新聞グループホールディングス 顧問）

沢田 登志子（一般社団法人ECネットワーク 理事）

前嶋 和弘（上智大学 教授、総合グローバル学部長）

森 亮二（英知法律事務所 弁護士）

柳川 範之（東京大学大学院経済学研究科・経済学部 教授）

山口 真一（国際大学グローバル・コミュニケーション・センター 准教授）

提言書の主要項目（今後の取り組みとして期待すること）

AIの積極的活用

誹謗中傷投稿抑止・削減のための取組み

措置ユーザーへの対応

- **AIや機械学習の積極的活用（＋活用拡大に伴う過剰削除等への手当）**
- 繰り返される違反投稿抑止に向けた施策

環境整備に向けた取組み

- 優良な投稿を奨励する取り組み
- ユーザーの選択に応じたコンテンツモデレーション

AIのアルゴリズムや生成過程の説明

透明化について

透明性レポート

ポリシー・削除基準の透明化

- **透明性レポートの策定**
- **ポリシーや基準、個別措置の理由の透明化、AIの合理性担保する制度設計など**

提言書の主要項目

	検討会での意見	提言
1 ポリシーと削除基準	■ ユーザーにとってどのような投稿が違反となるのか、禁止行為が分かりづらい	■ **ポリシーを明確化・透明化**すべき ■ **具体的な削除基準や削除例を公開**してはどうか
2 AI/機械学習	■ 利便性を損なうことなく、素早く、効率的に対応する必要がある ■ もっとも判断のブラックボックス化	■ **AIや機械学習**を用いた対応策の導入、拡大の検討が必要 ■ **AIのアルゴリズムの透明化**や合理性担保のためのガバナンス体制について説明
3 措置ユーザーへの対応	■ 過剰削除や誤削除の恐れがある ■ クレーム処理体制が整備されているか	■ **個別の措置理由を透明化**すべき ■ 不当に削除されたと主張する者からの**救済申立てフローの整備**が必要
4 透明性レポート	■ プラットフォーム事業者の社会的責任として取組みに関する社会への説明必要 ■ 社外からの監査を可能とし、さらなる改善を目指す	■ **自主削除件数や削除申請数、削除対応に当たる社内体制等を公開**すべき ■ **一定期間ごとに更新**すべき

提言に対するヤフーの対応方針

誹謗中傷対策におけるエコシステムの構築と透明化

1 ポリシーと削除基準

- **知恵袋は12/7に利用ルールの見直しをリリース**
- **Y!ニュースも削除例の追加等、随時見直し**

2 AI/機械学習

- **さらなる積極的な活用と精度の向上**
- **知恵袋でも本格導入検討開始**
- **アルゴリズムの内容や生成方法、合理性を担保するガバナンス体制等の透明化**

3 措置ユーザーへの対応

- **削除に関する問合せ窓口の設置**
- **措置理由の開示フロー検討**

4 透明性レポート

- **定期レポートの作成と公開（来春～）**
- **誹謗中傷対策に関する特設ページの設置**

5 繰り返される違反投稿抑止に向けた施策

6 優良な投稿を奨励する取り組み

7 ユーザーの選択に応じたコンテンツモデレーション

効果測定

フィードバック
対策の見直し

■外部有識者
■ユーザー
■社会全体
etc.

ヤフーの対応方針

ヤフーの対応方針に関する取り組みの実施状況

- 知恵袋におけるポリシーの見直し及び削除例の追加
  （ニュースにおいても削除例の追加検討）
- 繰り返す違反投稿抑止の施策
- AIを用いた反復違反投稿者への警告メッセージ掲出
- 削除に関する専用窓口の設置準備
- 2020年度版透明性レポートを本年夏までをめどに公表

など