

Yahoo!ニュースにおけるAIを活用した投稿時注意メッセージの掲出効果について

～ 不適切なコメントを繰り返し投稿するアカウントが13.5%減少 ～

ヤフー株式会社が提供する日本最大級のインターネットニュース配信サービス「Yahoo!ニュース」が、2020年7月からAIを活用して取り組んでいる投稿時の注意メッセージ掲出の効果により、「Yahoo!ニュース コメント」（以下、コメント欄）に不適切なコメントを繰り返し投稿しているアカウントが、掲出直後と比較して13.5%減少しました。



「Yahoo!ニュース」は、コメント欄における健全な言論空間を創出するため、専門チームによる人的なパトロールに加えて、自社で開発した「深層学習を用いた自然言語処理モデル（AI）」を利用して、誹謗中傷をはじめとした不適切な投稿を1日平均約2万件削除するなど、さまざまな対策を行っています。

これらの対策に加えて、「Yahoo!ニュース」は2020年より、不適切な投稿の事前抑止にも注力しています。同年7月には、パトロールに活用しているAI判定モデルを用いて、“不適切である可能性が高い”と判定されたコメントを過去数日以内に複数回投稿しているユーザーに対し、注意メッセージを掲出する取り組みを開始しました（※1）（図1）。これにより、過去の投稿内容やこれから投稿するコメントに乱暴な言葉づかいや他の人を傷つける表現がないかなど、ユーザー自身にあらためて考えるきっかけを提供することが本取り組みの目的です。

本取り組み開始直後の同年8月と比較して、4カ月後の同年12月には、コメントを投稿したユーザー数が増加したのに対して、注意メッセージが掲出されたアカウント数は13.5%減少しました（図2）。この結果から、本取り組みによって、不適切な投稿を繰り返すアカウントの減少に一定の効果が出ていることがうかがえます。

「Yahoo!ニュース」は今後も、コメント欄で投稿される多様な考えや意見によって、ユーザーがニュースに対する興味や多角的な視点を持つきっかけを提供するとともに、健全な言論空間を構築するために努めていきます。



abc*****

コメントを入力してください。

0/400

最近の投稿に他の人を傷つけるおそれのあるコメントが含まれると判定されました。心当たりはありませんか？ 乱暴な言葉づかいや他の人が傷つく内容がないか考えてみましょう。

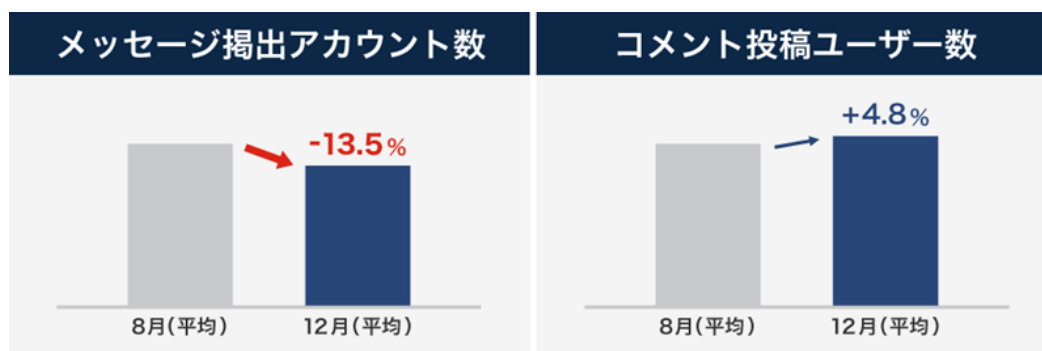
コメントポリシーと投稿内容を確認しましたか？

☐ はい、確認しました

投稿する

【図1】

不適切な投稿を繰り返しているユーザーに掲出される、投稿時の注意メッセージ



【図2】

左のグラフ：注意メッセージが掲出されたアカウント数（2020年8月と12月の比較）

右のグラフ：コメント欄における投稿ユーザー総数の日次平均

※1：「AIを活用した投稿時注意メッセージの掲出を開始しました」（2020年7月7日掲載、「Yahoo!ニュース」のオウンドメディア「newsHACK」）

<参考情報>

- ・ [Yahoo! JAPAN、個人に対する誹謗中傷などを内容とする投稿への対応や今後の方針を策定](#)（2020年12月23日発表）
- ・ [Yahoo!ニュース、違反コメント投稿者に対する投稿停止措置を強化](#)（2020年10月14日発表）
- ・ [Yahoo!ニュース、不適切コメントへの対策とし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APIを無償提供開始](#)（2020年9月18日発表）
- ・ [Yahoo!ニュース コメントの健全化に向けた取り組み](#)（2020年6月24日掲載、コーポレートブログ）
- ・ [Yahoo!ニュース コメントの取り組みまとめーコメント欄を「安心・安全な場所」にするためにー](#)（随時更新）