

メディア透明性レポート（2020年度版）



メディア透明性レポート作成の理由

ヤフーでは多数の投稿型プラットフォームサービスを提供しており、多くのユーザーに手軽に利用できる場を提供するとともに、安心して利用できるよう様々な施策を実施しております。

特に不適切な投稿への対応については、これまでも利用規約における禁止行為の明示やユーザーへの周知・啓発、専門チームによるパトロール、深層学習を用いた自然言語処理モデル（AI）を利用した対策などを実施してまいりましたが、社会情勢や外部の声も取り入れ、改めて当社の取り組みの見直し、強化、そして業界全体の健全化を図っていく必要があると考えています。

2020年には、プラットフォーム事業者としてユーザーに安心してご利用いただけるサービスを提供するために必要な自主ルールの策定を目的とした、「プラットフォームサービスの運営の在り方検討会」を設置し、さまざまな分野の専門家の方からご意見をいただいております。

同年12月には、この検討会より、個人に対する誹謗中傷などを内容とする投稿への対応について、プラットフォーム事業者による自主的取り組みの見直しと強化、取り組みの透明化などを促す提言書を受領し、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しています。

本件メディア透明性レポートは、同検討会の議論や提言書に基づき策定されたものであり、公開されたレポートについては、ユーザーや外部有識者からご意見を頂戴し、取り組みの再検証を行う等、自主的取り組みの継続的な改善のためのエコシステムの構築につなげてまいりたいと考えております。

なお、本件透明性レポートにおいては、ヤフーにおける主要な投稿型プラットフォームサービスである、

- ・利用者がニュース記事に関するコメントを書き込むことができる「Yahoo!ニュース」および
 - ・利用者相互が質問と回答を寄せ合い疑問を解決する「Yahoo!知恵袋」
- を対象に取りまとめを行っております。

プラットフォーム事業者の社会的責任を果たすべく、今後も対応の一層の強化を図ってまいります。

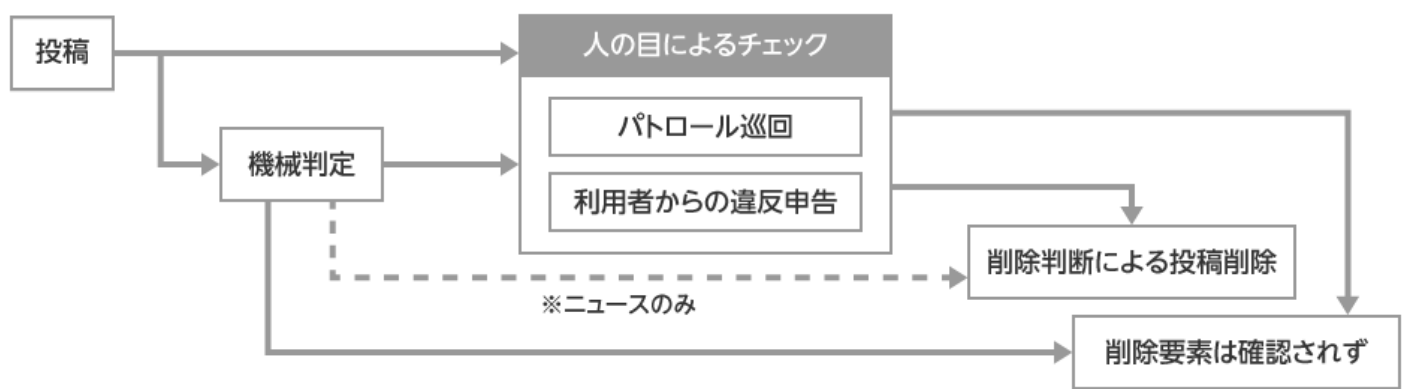
投稿の監視体制

ニュース・知恵袋いずれも、当社が独自に開発したディープラーニングに特化したスパコン「kukai」を活用した機械学習システムを導入し、迅速に利用規約・ガイドラインに抵触する投稿が検知できる仕組みを強化しています。

ユーザーによる投稿後、当該投稿がガイドラインに抵触する内容である確率を判定し、明確にガイドラインに抵触している投稿は自動的に削除がなされ、明確とは言えないが、抵触している可能性がある投稿は、優先的にパトロール部隊による目視確認による検討フローに移され、より正確な判断を行うよう二重チェックの体制がとられています。

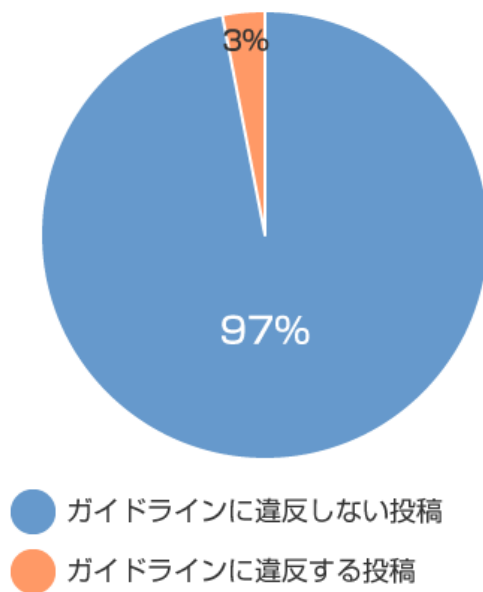
また、目視確認として、kukaiによる判断結果のチェックのほか、利用者よりご連絡いただく違反投稿報告のチェックおよび投稿の自主的なパトロールも並行して定常的に行われています。

※知恵袋においては2021年12月現在、機械判定により自動的に削除されるフローはありません。

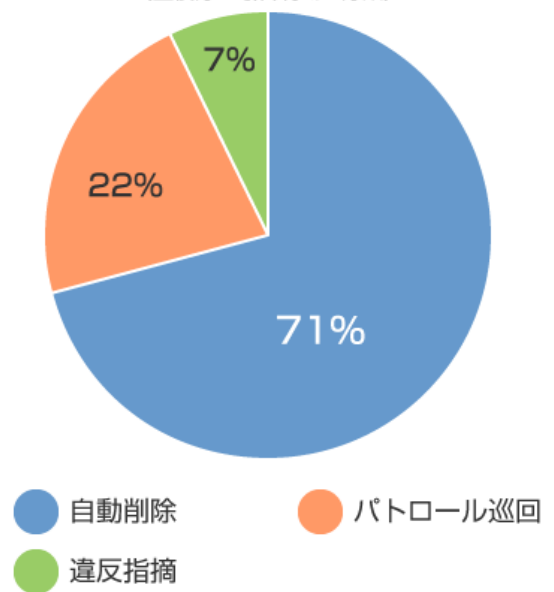


■ ニュース

総投稿数に対しての削除数

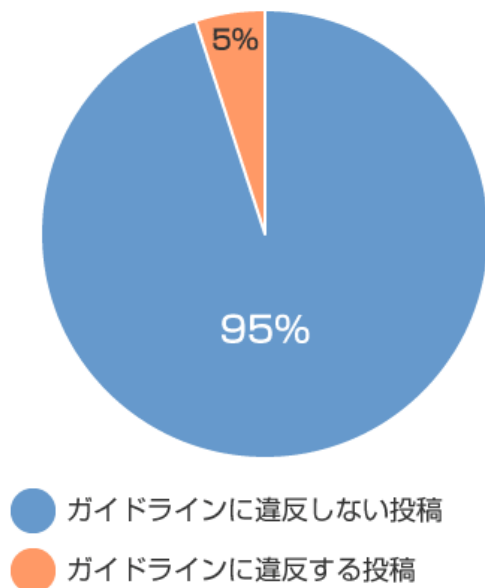


自動削除／人的チェックの内訳
(重複分は最終判断で分類)

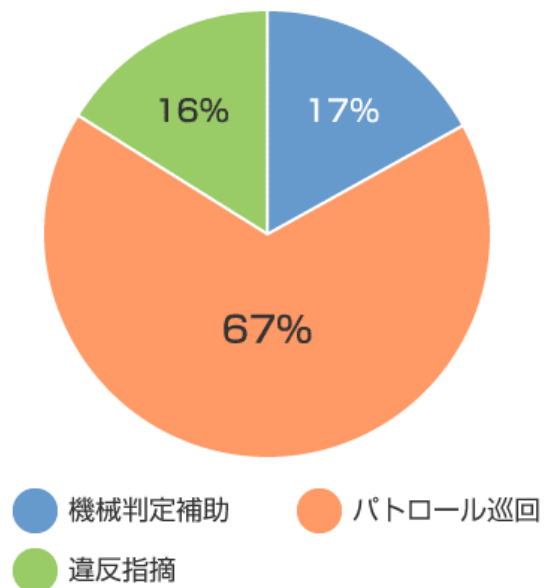


■ 知恵袋

総投稿数に対しての削除数



自動削除／人的チェックの内訳
(重複分は最終判断で分類)



※ 2021年3月の数値です。

(2022.12.6追記)

知恵袋における削除数について、機械判定数の一部を誤って削除数に計上していたことが判明したため、訂正を行っています。

関係者の皆様には、深くお詫び申し上げます。

(誤) 約30万件/月 ⇒ (正) 約23万件/月

パトロール体制について

パトロール体制

- ・ 24時間365日稼働
- ・ 人員数：約70人

※投稿の削除等の監視業務を専門とする人員数

ヤフーでは日々大量になされる投稿への監視強化のため、システムによる全投稿のチェック体制の整備や、疑わしい投稿を人の目で優先的に確認できるような仕組みづくりに努めています。

教育体制

- ・利用規約やサービスごとに定めるガイドライン等の規定の他、内部の運用マニュアルを作成し、担当者ごとに判断にぶれが生じることのないように努めています。
- ・配属時に1ヶ月以上の基礎研修（座学、モニタリング、OJT）を受けた専門スタッフが対応しています。
- ・ガイドラインに抵触しているか否かについて疑義が生じた投稿についてはチーム内での協議を経て複眼的な判断がなされるほか、各サービスの担当者や法的判断を行う専門部門へのエスカレーションを実施し、適切な判断がなされるよう体制構築をしております。

措置の内容

- ・対象投稿の削除
- ・違反の重大性や回数に応じた利用停止措置
- ・Yahoo! JAPAN ID登録情報である携帯電話番号を削除履歴と照合しブロック
（参考）「[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](#)」

なお、機械的な対応として、サービスによっては、過去に違反投稿を複数回行ったユーザーに対する警告表示や、投稿の非公開措置（投稿をクリックしなければ全文を閲覧することができなくなる措置）を実施しています。

機械的な対応について

ニュースにおけるAIや機械学習を利用した対策

ニュースコメントでは、建設的モデル、関連度モデル、不適切投稿判定モデル（※）という3種類のモデルを使用しています。

※「不適切投稿判定モデル」は、個人に対する誹謗中傷などを内容とする投稿への対応をテーマとした「プラットフォームサービスの運営の在り方検討会」の提言書において、「不快モデル」として記載していたものです。

- ・各モデルについて

不適切投稿判定モデルは、過度な批判や誹謗中傷、差別、わいせつや暴力的などのガイドラインの項目に違反するおそれのあるコメントをAIが点数化するモデルとなっています。同モデルで対象となるコメントは、「暴力的、差別的、過度に品位に欠けるコメント」等であり、同モデルは、算出されたコメントの点数をもとに、コメントの掲載順位の変更や、自動的な削除などに活

用されます。

建設的モデルは、建設的と考えられる良いコメントの順位を上部に表示させるためのモデルであり、「自分の意見を元に議論を喚起する発言」「客観的で、必要であれば根拠を提示している発言」といった評価軸でスコアを算出するものです。

関連度モデルは、ニュース記事とコメントの関連度をAIが判定し、関連度が低いと判定されたコメントは自動削除、あるいはパトロール部隊による目視による確認フローに移され、パトロールの効率化に役立てられています。

・各モデルの生成方法について

不適切投稿判定モデルと関連度判定モデルの生成方法については、専門チームがパトロールを行った結果を正解データとして学習し、その判断を再現します。学習の際には、Yahoo! JAPANが持つ大規模テキストデータを言語理解力向上のためにあわせて活用しています。

性能とプロセスの検証を実施し、最新技術を取り入れて、継続的にモデルを改善しています。

建設的モデルの生成方法は、一定数の一般ユーザーの集団に対し、「建設的コメントとなる条件」の定義を示し、「建設的コメント」を選択できなかったユーザーのみ対象から除外したうえで、記事とコメントのセット（1つの記事に対しコメントが複数個あり）を与え、建設的か否かの判断を行ってもらい、その正解データを利用して生成しています。また、モデル生成後も、効果検証を実施し、導入前と比して、社内担当者の目から見ても適切な並び順となっているか、コメント欄の閲覧者が増えているのか等を参考に、定量的な指標をベースとして、効果の有無を測定しています。

・kukaiの活用について

2020年3月、関連度モデルについて導入をされていた最先端の自然言語処理モデルとスーパーコンピュータ「kukai」を活用したコメント対策の対象を、暴力的、差別的、過度に品位に欠ける等のコメントに展開し、誹謗中傷投稿へ対策が強化され、違反コメントの検知件数が約 2.2 倍に向上しました。

また、2021年からは、建設的モデルを利用したAIの外部への無償提供を開始し、複数の外部投稿系サービス事業者様に導入いただいています。

知恵袋における機械学習、AIを用いた対策

知恵袋では、機械判定に用いられるモデルとして、低品質投稿判定モデルがあります。低品質投稿とは、アダルト関連や文意不明の（質問/回答になっていない）投稿、誹謗中傷を含む一般人が不快と感じる投稿を指し、過去のパトロール人員が低品質投稿と判断した投稿を学習データと

して利用し、機械判定モデルを生成しています。低品質投稿のスコアを算出し、一定の閾値以下の投稿については、自動的に回答の非公開措置（投稿をクリックしなければ全文を閲覧することができなくなる措置）を行うほか、優先的なパトロールチェック、ランキング表示からの除外などの措置が行われています。

基準のアップデートを定期的に行っており、（直近では2020年夏に実施）、効果検証として、アップデートの前後における低品質投稿の対応件数を計測しています。当該データによれば、低品質判定がなされた投稿件数が増加しており、低品質判定能力の向上及びユーザーが目にする低品質投稿数の減少が確認できております。

Yahoo!ニュース、Yahoo!知恵袋の利用状況

■ ニュース

- ・ 投稿数：約1,050万件/月
- ・ 削除数：約35万件/月

■ 知恵袋

- ・ 投稿数：約450万件/月
- ・ 削除数：約23万件/月

※ 2021年3月の数値です。

（2022.12.6追記）

知恵袋における削除数について、機械判定数の一部を誤って削除数に計上していたことが判明したため、訂正を行っています。

関係者の皆様には、深くお詫び申し上げます。

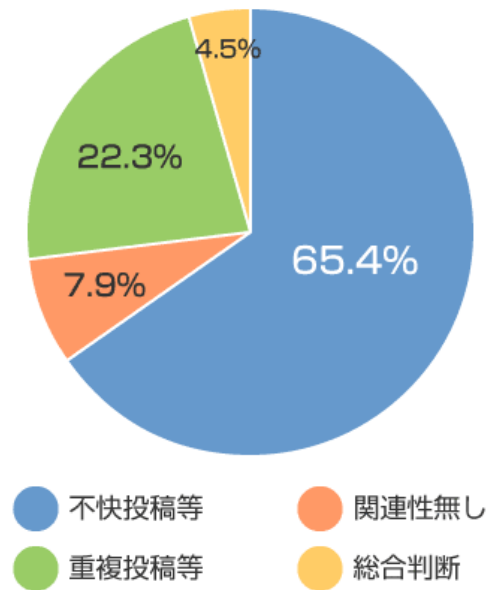
（誤）約30万件/月 ⇒ （正）約23万件/月

削除理由の内訳

■ ニュース

＜削除理由の内訳＞（注1）

削除理由の内訳



※注1 人の目によるチェックを経た削除理由の内訳

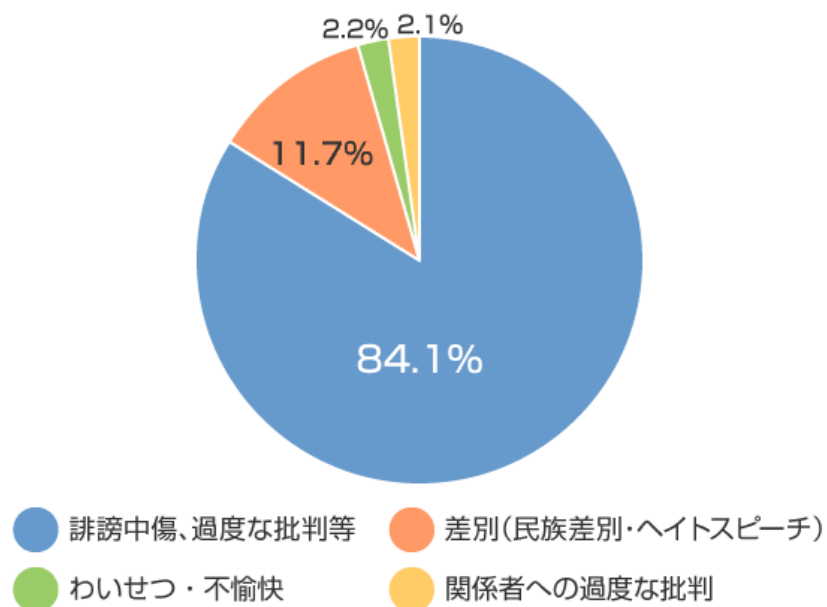
※ 「不快投稿等」には、以下の投稿が含まれます。

- ・過度な批判や誹謗（ひぼう）中傷、不快な内容
- ・わいせつや暴力的等、不愉快な内容

※ 「総合判断」には、個々のガイドラインに違反するものではないが、総合判断として不適切と判断された投稿が対象

<不快投稿等の削除理由の内訳>（注2）

不快投稿等の削除理由の内訳



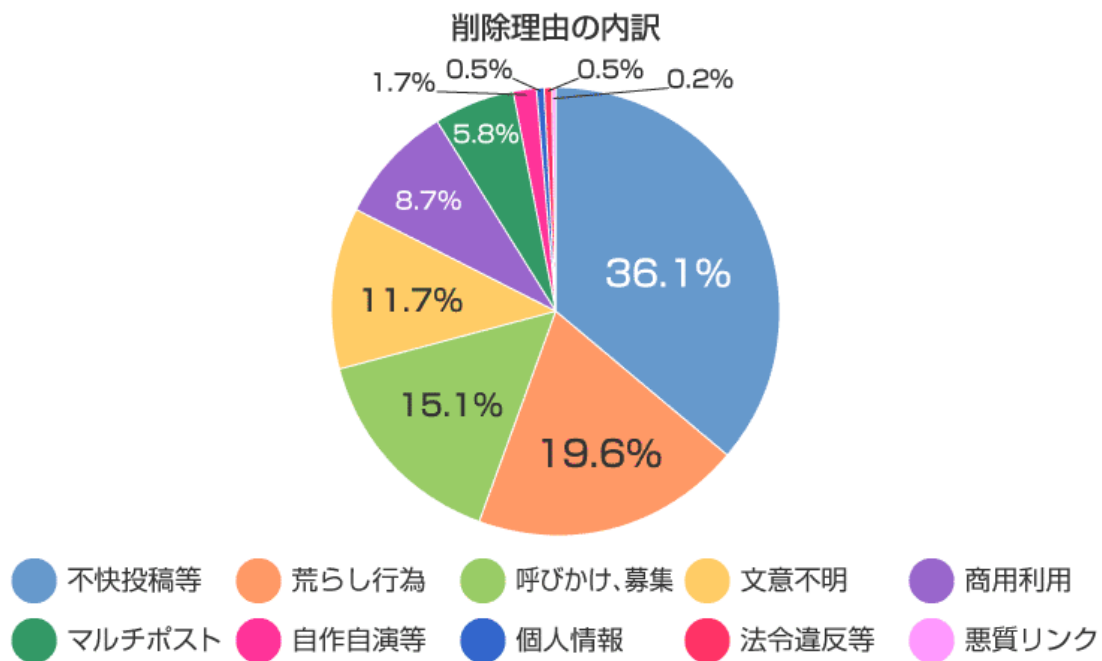
※注2 全体の5%をサンプリング調査

※ 「誹謗中傷等」には、誹謗中傷、過度な批判、攻撃的な投稿などが含まれます。

- ※ 「差別」には、特定の民族や国に対して、危害を加える旨の告知や攻撃、暴力的表現、著しい侮蔑、社会からの排斥などの投稿を含みます。
- ※ 「関係者への過度な批判」には、事件関係者への過度な批判が含まれます。

■ 知恵袋

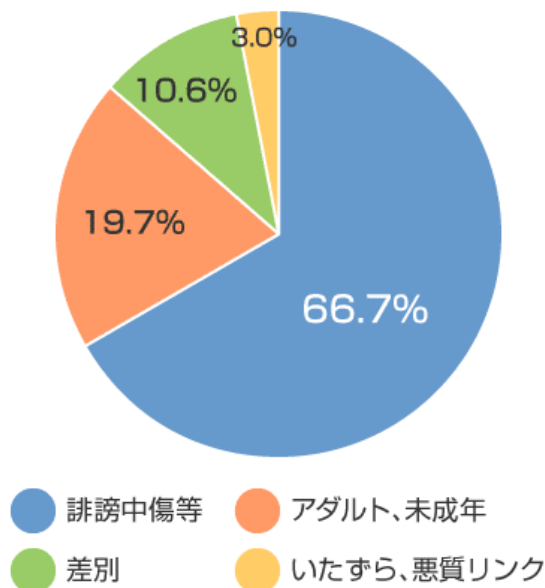
<削除理由の内訳>



- ※ 「個人情報」とは、個人を特定できる情報を含む投稿を指します。
- ※ 「法令違反等」とは、法令違反行為や犯罪行為を誘発する投稿を含みます。
- ※ 「自作自演等」とは、他人になりすましたり、自作自演行為によって、質問、回答を繰り返す投稿を指します。
- ※ 「呼びかけ、募集」には、SNSなどへの招待依頼、他の知恵袋への回答依頼など、他人やほかの利用者への勧誘や呼びかけ行為が含まれます。

<不快投稿等の削除理由の内訳>（注3）

不快投稿等の削除理由の内訳



※注3 全体の1%をサンプリング調査

※ 「誹謗中傷等」には、誹謗中傷、過度な批判、攻撃的な投稿などが含まれます。

※ 「差別」には、特定の民族や国に対して、危害を加える旨の告知や攻撃、暴力的表現、著しい侮蔑、社会からの排斥などの投稿を含みます。

プロバイダ責任制限法に基づく対応状況

対応体制

利用規約・ガイドラインに抵触する投稿は、「投稿の監視体制」に記載したとおり、機械学習システムによる自動削除、あるいは自主パトロールや利用者からの違反報告をきっかけとする削除が行われています。

また、他ユーザーの投稿等によってご自身の権利を侵害されたという理由に基づき、プロバイダ責任制限法に関する申告（当該投稿等の削除や発信者情報の開示）を、裁判外の書面にて行うことが可能です。

[プロバイダ責任制限法に関する申告を行う方へ](#)

プロバイダ責任制限法に関する申告を受けた場合、法的判断を行う専門部門にて、プロバイダ責任制限法等の法令を遵守し対応を行っています。専門部門においては7～8名程度で対応し、判断に悩ましい投稿については、一担当者だけの判断によらず、複数人での協議を経て判断しています。そのほか社外弁護士複数名に諮問して対応することもあります。また、削除請求について

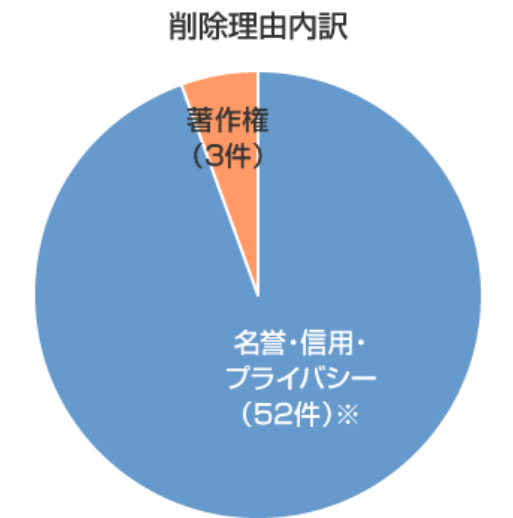
は必要に応じて、開示請求についてはプロバイダ責任制限法に従い原則として全件、投稿者に意見照会を行っています。

さらに裁判手続によってこれらの請求を受けることもあります。この場合も法令を遵守し速やかに対応しています。

対応状況

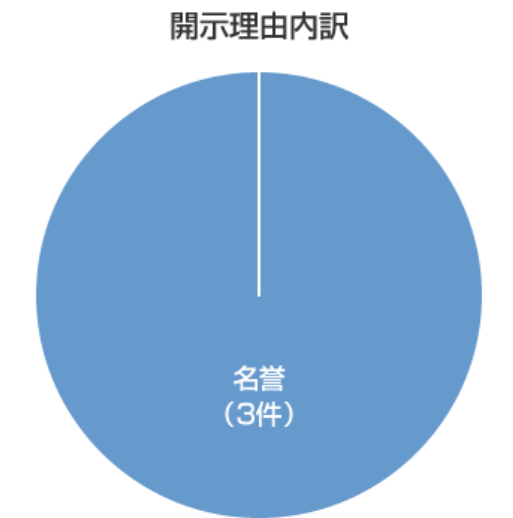
2020年度の対応状況は下記の通りです。

■ 裁判外の書面による請求（注1）



参考：削除請求件数 118 件
(請求理由内訳：名誉・信用・プライバシー 115 件、著作権 3 件)

※うち 5 件は行政機関からの削除請求によるもの
(行政機関からの削除請求件数 11 件)

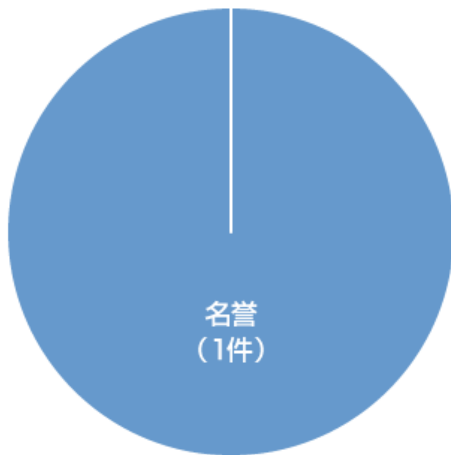


参考：開示請求件数 28 件
(請求理由内訳：名誉・信用・プライバシー 26 件、著作権 2 件)

- ・ 件数とは、投稿数ではなく書面数をいう（つまり、1件につき複数投稿に関する請求が含まれている場合がある）
- ・ 削除/開示した件数には、一部の投稿のみ削除/開示した場合も含む

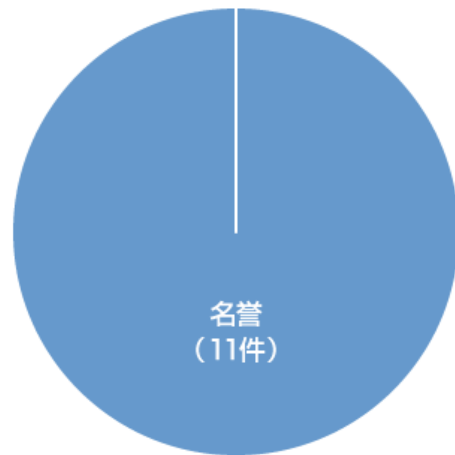
■ 裁判上の請求（注2）

削除理由内訳



参考：削除請求件数 6 件
(請求理由内訳：名誉 5 件、プライバシー 1 件)

開示理由内訳



参考：開示請求件数 20 件
(請求理由内訳：名誉 19 件、プライバシー 1 件)

- ・件数とは、投稿数ではなく事件数をいう（つまり、1件につき複数投稿に関する請求が含まれている場合がある）
- ・削除/開示した件数には、一部の投稿のみ削除/開示した場合も含む
- ・一部削除/開示には、裁判所が却下心証を示したために債権者が一部の投稿に関する請求を取り下げた件数も含む
- ・削除/開示には和解による削除や任意削除/開示も含む
- ・同一案件であっても、審級等が異なる場合は別件とする

※注1 2020年度に請求を受けたもののうち、2021年6月30日時点で結論が出ているもの

※注2 呼出状の日付が2020年度のものうち、2021年6月30日時点で結果が出ているもの

措置ユーザーへの対応

知恵袋：削除された質問・回答については、My知恵袋内の［取り消し・削除］のタブから削除理由を確認できます。削除理由が公開されているのは、削除された日から2週間です。

ニュース：問合せ窓口にて個別にお問い合わせください。削除理由について確認させていただきます。