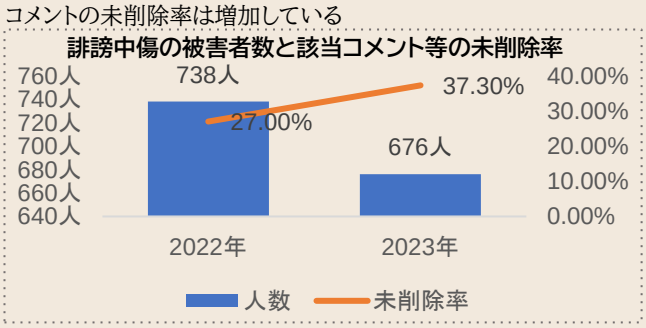
ChatGPTで人権を守ることができるか

研究背景・目的

・SNSでの誹謗中傷によって自ら命を経ってしまった人がいることをニュースで知った
・実際，誹謗中傷による被害者は減少しているが，削除申請をしたコメントの未削除率は増加している



・全員の人権を守るにはコメント投稿者の「表現の自由」を制限してはいけない
→ 現状，人で判断が難しい「誹謗中傷」と「批判」について，「誹謗中傷」と「批判」の判別できるAIを開発したい

※グラフは一般社団法人セーファーインターネット協会誹謗中傷ホットライン 活動報告（2022年1月1日～2022年6月30日）誹謗中傷ホットライン 活動報告 （2023年1月1日～2023年6月30日）より作成

研究対象

容姿等に対する誹謗中傷が多いと考えられる美容系YouTuberが投稿した動画を対象とした
動画再生回数は10万回以上、コメント数は100件以上とする

研究方法

1本の動画に対する全てのコメントを収集する
協力者に提供していただいたプログラムを活用し，APIを用いて合計49本のコメントを分析する

↓

ChatGPTにコメントを入力する
項目ごとに0～10の整数で判別した結果を出力する

↓

ポジティブコメントとネガティブコメントの割合を分析する

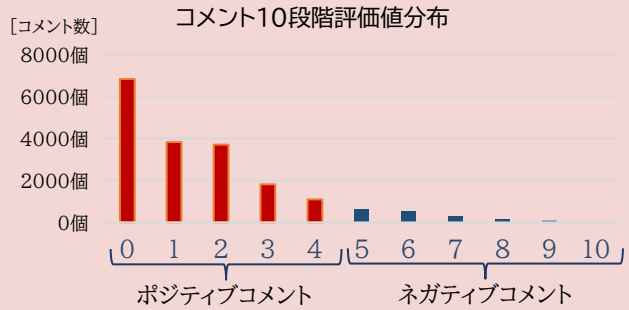
↓

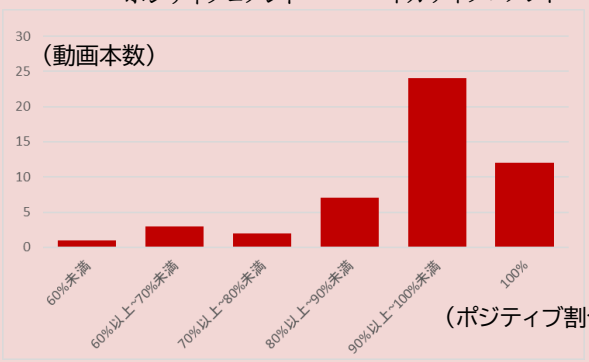
ネガティブコメントについて，詳細に解析する

分析1　動画全体のポジ・ネガ割合の把握

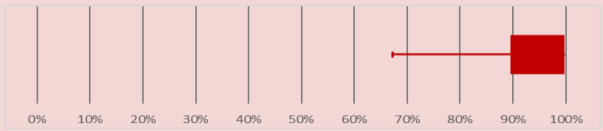
2023年9月13日～9月24日における「美容系」と検索したときの**上位49本の動画**を対象に調査を行った

ネガティブ度の判断は，ChatGPTが0～10と判断した数値において，0～4をポジティブコメント(P)，5～10をネガティブコメント(N)とした．





1つの動画のポジティブコメント割合の本数



1つの動画のポジティブコメント割合の箱ひげ図

・49本の動画中ポジティブコメントの方が多い
・ネガティブコメントには誹謗中傷コメント以外に批判コメント、動画投稿者を心配するコメント等が含まれる

分析2　ネガティブコメント分析

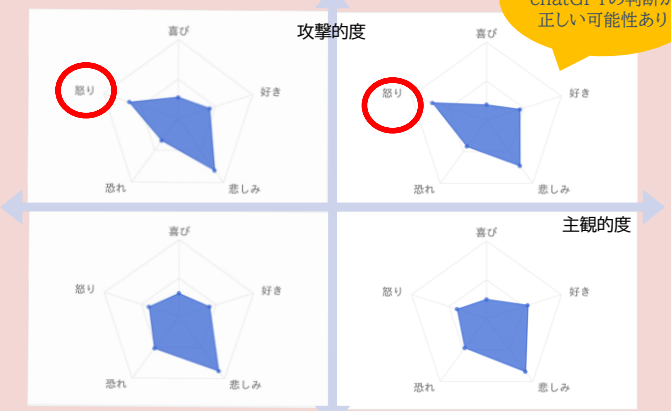
ChatGPTが出力した「誹謗中傷」と「批判」の相違点に基づき，ネガティブコメントを①攻撃的度②主観的度によって分別した。

出現頻度順ワードクラウド

内容に関係する単語が多い

感情分析

感情分析結果よりchatGPTの判断が正しい可能性あり



※ユーザーローカルAIテキストマイニングによる分析(https://textmining.userlocal.jp/)

考察

・ポジティブコメントについて，動画の内容に即した単語が多く、使われた単語だけでは「誹謗中傷」と「批判」のコメントの識別は難しい
・成果として，**『攻撃的であるかどうか』**は感情分析から**判断可能**であることがわかった

まとめ・今後の展望

ChatGPTを活用し，『攻撃的なコメント』の判別可能

➡ 攻撃的と判断したコメントから，「誹謗中傷」と「批判」を選別する特徴を見抜くAI開発を行う

協力　一般社団法人デジタル人材共創連盟　一般社団法人 i-RooBO Network Forum　猪熊祐斗