SecHack365

攻撃的投稿調査ツール

Tesave

思索駆動コース 大畑和也

Background

ネット上の誹謗・中傷が社会的に問題化している。
特にSNSにおいて著名人を対象にした攻撃的な投稿は
被害者の名誉を傷つけ、精神的苦痛を負わせる事件として
報道されることもあった。

情報セキュリティ10大脅威 2021 (個人)

順位	項目
1位	スマホ決済の不正利用
2位	フィッシングによる個人情報等の詐取
3位	ネット上の誹謗・中傷・デマ

IPA 情報セキュリティ10大脅威 2021 より一部抜粋

被害者が置かれている状況を周りが把握し
適切なアクションを行う事が必要だが、
個別の投稿や全体の様子を把握するために
膨大な投稿の確認は困難である。

代表的なSNSであるTwitterにて特定の人物を対象にした攻撃的投稿を調査するシステムを開発した。

Method

攻撃的表現の検知

攻撃単語の抽出イメージ

69000件以上の投稿を収集し機械学習モデルに適用させるためのデータセットを独自に作成した。
攻撃的な表現が特に多い掲示板から投稿を収集。
データは各投稿ごとに分かち書きして単語に分割した後、ベクトルで表現した。

400語の攻撃性が疑われる単語を、報道や書籍等で取り上げられた攻撃的単語の関連語として抽出した。
自分たちが認知していない攻撃的単語を抽出できた。

150語以上の攻撃性が特に高い単語を選別して利用した。
攻撃性はポジティブ/ネガティブ単語群と共に利用されている割合をネット上の使用頻度から算出した。

抽出した関連単語と攻撃スコアの一覧。
攻撃的な単語を抽出することができ
そのスコアも攻撃性が高い結果となった。

「アホ」の関連語とそのスコア

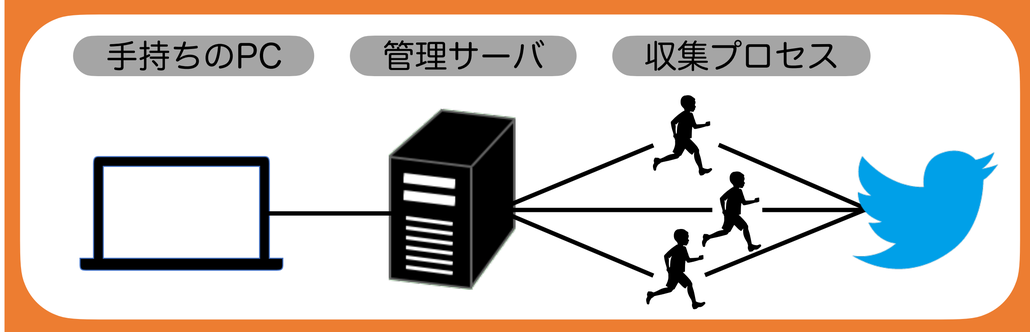
順位	単語	攻撃スコア
1位	バカ	2.2
2位	カス	3.0
3位	馬鹿	3.0

社会的な影響力の考慮

影響力が強い投稿は攻撃的投稿を増やすきっかけに繋がる場合がある。
投稿の言語的な解析では難しい、社会的な影響力及び拡散力についてフォロワーの数等を利用し、最終的な投稿の危険度を算出した。影響力の強い人物の攻撃的投稿はスコアが高くなりやすく、認知しやすい構造となっている。

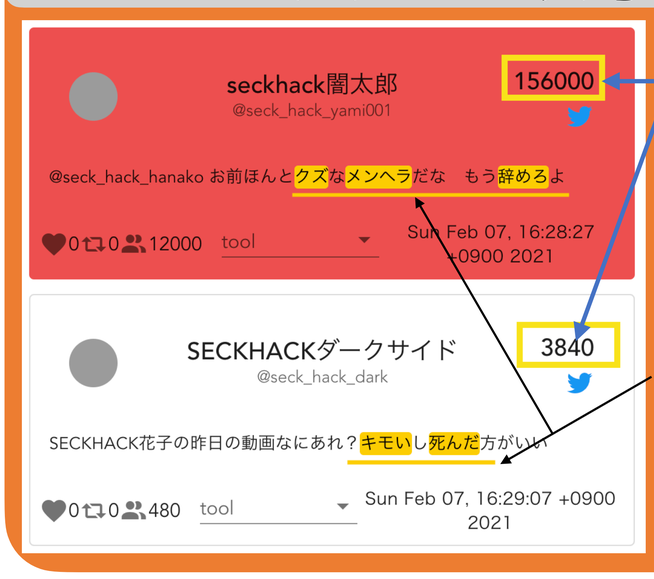
Product

Tesaveは各自で管理サーバを準備し、その上に構築するシステム。管理サーバはTwitter上の投稿収集及び取得データの保存を行う収集プロセスを管理する他、データベースやWebサーバの機能が集約されている。利用の際には手持ちのPCからブラウザを開くだけで投稿の収集・調査が行えとても簡単。



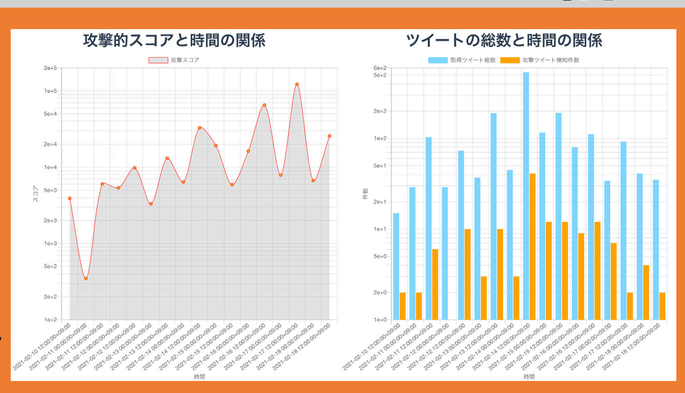
- 40人を超える対象者の同時並行による収集を独立した収集プロセスの生成により実現した。複数人の調査が容易に行える
- 対象となる人物が言及されている投稿や、直接のリプライについての投稿を収集する。過去の遡り、最新の投稿取得について選択できる
- 活躍する著名人のほか、これからデビューするアーティストやモデル動画投稿者など次世代を拓く人たちへの支援にも有効

「あの人への攻撃が心配」それなら Tesave で詳しく見てみよう！！



危険度が高い投稿を優先的に表示する。
攻撃的表現が強く、影響力の大きい注意すべき投稿から確認することができる。

検知した攻撃的表現を瞬時に視認できる。
本文を全部読まずとも検知した攻撃表現がどこに存在するかがわかる。



横軸には時間を、縦軸に合計スコアや投稿の件数を設定したグラフを表示。
炎上検知・鎮静化の確認に役に立つ。

単語編集ページでは単語の追加や削除、ウエイトの変更が自由に行える。
対象者の状況に合わせた調査が可能。

Future work

- 芸能事務所・マネージャーの方に使用していただき効果を確認
- Twitter以外のSNSへの拡大

国立研究開発法人 情報通信研究機構

連絡先: shiroichigodev@gmail.com