ChatGPTを用いたネット炎上を防ぐセルフ検閲アプリの開発

総合情報学科 知能情報システム研究室
永井ゼミ 4年
横山　翔太

課題と背景

- ネット炎上は増え続けている
  - ネット炎上は2006年で約41件,2010年で102件,2015年で1002件[1]、2022年では1570件[2]発生したとされる
  - ここでネット炎上は「ある人物や企業が発信した内容や行った行為について、ソーシャルメディアに批判的なコメントが殺到する現象」(山口,2016)とする
- 炎上は個人や企業に大きな影響を与えるので、対策が必要
  - 所謂バイトテロによる金銭的被害
  - 学校・企業からの退学・解雇
  - 企業の営業悪化、あるいは株価下落

課題と背景

- 私が高校の頃に取り組んだ「感情を読み取るチャットアプリ」
  - インターネット上でのトラブル防止を目的
  - テキストのネガポジを判定→事前に表示
  - 形態素解析→極性辞書で判定

形態素解析とは

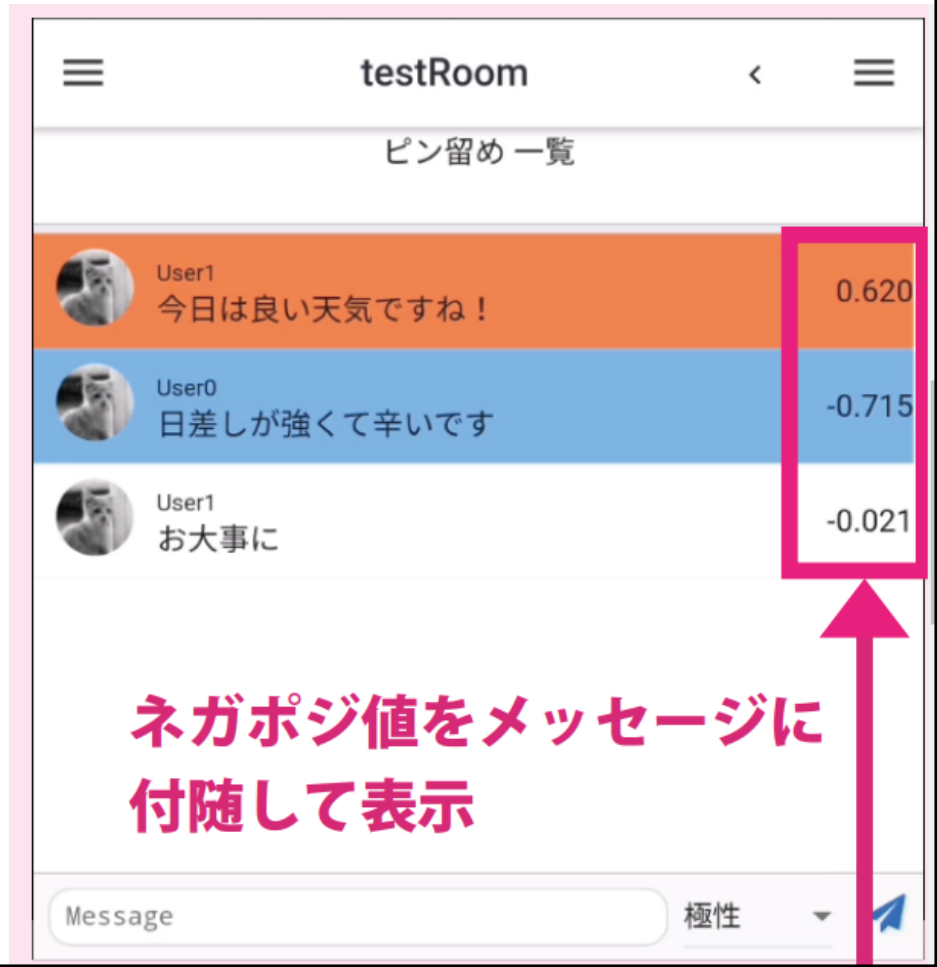
私/は/怒っ/て/いる

{"surface_form": "私",
"pos": "名詞",
"pos_detail_1": "代名詞",
"pos_detail_2": "一般",
"basic_form": "私",}

{"surface_form": "は",
"pos": "助詞",
"pos_detail_1": "係助詞",
"basic_form": "は",}

{"surface_form": "怒る",
"pos": "動詞",
"pos_detail_1": "自立",
"conjugated_type": "五段・ラ行",
"conjugated_form": "基本形",
"basic_form": "怒る",}

…

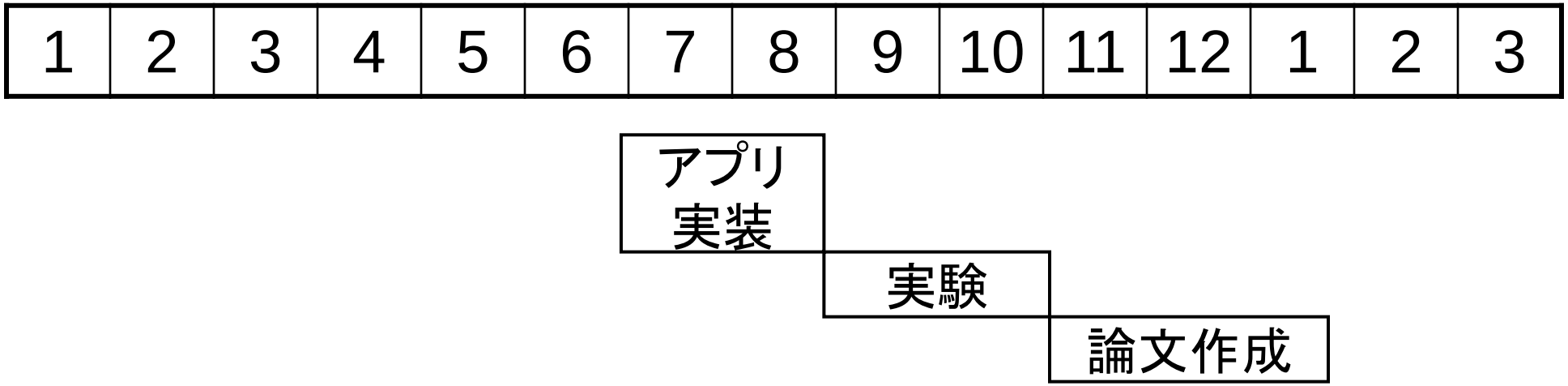


ChatGPTを用いたネット炎上を防ぐセルフ検閲アプリの開発

- 目的
  - 検閲アプリを作成する。テキストを入力すると、ChatGPTが文面をチェックし、内容や表現について炎上しそうなところを表示、大体表現を提案することでトラブルを未然に防ぐ
  - LLMのAPIを活用したアプリ開発のベストプラクティスの考察

開発環境と構成・研究計画

- クライアント&サーバのwebアプリ
  - フレームワーク：react(CRA or Next.js)
  - サーバ：node.js
  - API：Chat-GPT API
- 研究計画



プロンプト

- LLMで質の高い回答を得るために色々なテクニックが考案されている
  - 立場や役割を与える(職業など)
  - 与える指示を構造化する
  - 回答例を与える
  - 一度に複数の指示ではなく、ひとつずつ指示を与える

###指示###
あなたはコンプライアンスの専門家です。与えられたテキストからネット炎上しそうな語を抜き出してください。
次に、抜き出した語にネット炎上しそうな度合いを0～100点で点数をつけてください
###フォーマット###
JSON形式
###テキスト###
ホントクソな質問だよな。
お前死んだ方がいいよ。
肉の写真をアップしてるのはインスタで和牛を広めたいからってんの見てわからない？
お前クソ？
そもそもふつうに野菜は好きだからたくさん食べてるけどインスタでそれをアップしてもクソの役にも立たねーだろ。
それを野菜とかを子供の頃嫌いだっただろうお前みたいなやつらに邪智されたくないね。
おれは美味しい野菜を子供の頃からたくさん食べててむしろ肉より野菜の方が好きだった。
それは食ってた肉がまずかったからだ。
それを良薬口に苦し的な文脈で野菜を食べた方が健康ですよって上からマウントでクソコメント送るなボケ。
二度と来るな。

json　コードをコピーする

```
{
  "炎上語": ["クソ", "死んだ方がいい", "邪智", "マウント"],
  "炎上度": {
    "クソ": 90,
    "死んだ方がいい": 95,
    "邪智": 85,
    "マウント": 80
  }
}
```

検出すべき表現

- 山口(2016)では、炎上を①誰が、②何をしたか、③どう対応したか、で分類している。
- 右図の②何をしたか、における I と II を取り扱う。
- 特に I においては更に解釈を広げて「コンプライアンスに反する行為(の告白・予告)」として扱う。
  - コンプライアンスは本来は単に「法令遵守」を意味するが最近では暗黙的な社会規範や倫理観も内包してきている

番号	大分類	記号	小分類
①	誰か	A	著名人
		B	法人等
		C	一般人
②	何をしたか	I	反社会的行為や規則・規範に反した行為(の告白・予告)。
		II	何かを批判する、あるいは暴言を吐く(政治・宗教・ネット等に対して)。デリカシーのない発言をする。特定の層を不快にさせるような発言・行為をする。
		III	自作自演、ステルスマーケティング、捏造の露呈。
		IV	ファンを刺激(恋愛スキャンダル・特権の利用)。
		V	他者と誤解される。
③	対応	㋐	挑発、反論、主張をとおす。
		㋑	コメント削除。
		㋒	無視。
		㋓	謝罪、発言自体の削除、発言撤回の発表。

山口真一 ネット炎上の研究 「炎上の分類・事例と炎上参加者属性」(2016)より引用

参考文献

- [1]総務省.”情報通信白書 令和版”.総務省.2019-7. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r01/html/nd114300.html,(参照2024-7-4)
- [2]シエンプレ株式会社.”デジタルクライシス白書2023”.シエンプレ株式会社.2023-1-31. https://www.siemple.co.jp/document/hakusho2023/,(参照2024-7-4)
- 山口 真一.”ネット炎上の研究 「炎上の分類・事例と炎上参加者属性」”.国際大学グローバル・コミュニケーション・センター.2016.https://www.glocom.ac.jp/wp-content/uploads/2016/04/20160510_Yamaguchi.pdf,(参照2024-7-4)