2021日心ワークショップ・J-LIWC2015

辞書ベースのテキストマイニング
ML-Ask, J-LIWC, J-MDFを例として

笹原 和俊[1]、奥田 慎平[2]

1. 東京工業大学 環境・社会理工学院
2. 名古屋大学 大学院情報学研究科

発表内容

1. テキストマイニングとは
2. 辞書ベースのテキストマイニングの事例

発表内容

1. テキストマイニングとは
2. 辞書ベースのテキストマイニングの例

テキストマイニング

大量のテキストから自然言語処理技術を用いて有用な情報を抽出する

- 統計分析
  - 頻度分析、相関分析、回帰分析、感情分析、ネットワーク分析 等
- 機械学習を用いた分析
  - 潜在的意味解析（SVD）、単語埋め込みモデル（word2vec）、トピックモデル（LDA）など
  - 深層学習による自然言語処理（BERTなど）

テキストの例

- 自由回答のアンケート
- ソーシャルメディア（掲示板、SNS、ブログ、メーリングリスト）
- 論文の抄録
- 議事録、裁判記録
- 電子カルテ
- コールセンターに寄せられた苦情、顧客とオペレーターのやりとり
- 製品やサービスのレビュー（ECサイト、レストラン比較サイト）

テキストマイニングの実応用

- 多様な意見や潜在的ニーズの把握
- 設問にとらわれない自由な感想を把握
- 消費者からのリアルタイムな意見の獲得
- 業務の問題点の発見
- ナウキャスト
- トレンドや予兆の検出

テキストの前処理 ← テキストマイニングの99%はこの作業😭

1. クリーニング
2. 正規化：表現揺れの統一
3. 形態素解析：分かち書き、品詞付与
4. 基本形への変換：語幹処理、見出し語化
5. ストップワード（不要語）の除去
6. 単語の数値化

形態素解析

テキストデータを形態素（意味を持つ最小単位）に分割し、
それぞれの品詞等を判別する作業

- 友人とニューヨークの美術館に行った。

⬇ 形態素解析

- 友人　と　ニューヨーク　の　美術館　に　行っ　た　。

名詞　助詞　名詞　助詞　名詞　助詞　動詞　助動詞　記号

英語だとStanford POS Tagger、日本語だとMeCabが有名なツール

MeCabによる形態素解析の例

今日はいい天気だ。

```
今日	名詞,副詞可能,*,*,*,*,今日,キョウ,キョー
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
いい	形容詞,自立,*,*,形容詞・イイ,基本形,いい,イイ,イイ
天気	名詞,一般,*,*,*,*,天気,テンキ,テンキ
だ	助動詞,*,*,*,特殊・ダ,基本形,だ,ダ,ダ
EOS
```

📕MeCab

- オープンソースの日本語用の形態素解析器（https://taku910.github.io/mecab/）
- 品詞タグ付けに条件付き確率場（CRF）を利用

辞書ベースのテキストマイニング

- 概念辞書：WordNet
- 感情辞書：ML-Ask
- 心理学カテゴリー辞書：LIWC、J-LIWC
- 道徳基盤辞書：MFD、J-MFD

メリット

- 専門家が作った辞書は信頼性が高い（場合が多い）
- 結果の解釈が容易

デメリット

- 網羅性が高くない
- アップデートが頻繁でない（されない）
- 辞書の作成にコストがかかる

J-LIWC

- 心理学的カテゴリを計測するソフト（辞書）
- テキストを入力として、各英単語を言語学的カテゴリや心理学的プロセスに関するカテゴリにマッピングする (Pennebaker 2015)
  - 約6400単語、73カテゴリ
- J-LIWCはその日本語版
  - 約11,600語、69カテゴリ

J-LIWCのカテゴリと単語の例（五十嵐・笹原 2018）

知覚的プロセス	Percept	652	436	
視覚	See	190	126	光, 風景, 目
聴覚	Hear	142	93	歌, 笑, ひそひそ
感覚	Feel	206	128	硬い, 厚い, 鋭利
生体的プロセス	Bio	900	748	
身体	Body	262	215	筋肉, 胸, ヒップ
健康	Health	371	294	肥満, 免疫, 元気
性	Sexual	149	131	妊娠, 性器, セックス
摂取	Ingest	224	184	アルコール, 砂糖, 摂取
動因	Drives	1866	1103	
親和	Affiliation	395	248	愛*, コミュニティ, 談話
達成	Achieve	439	213	勤勉, パーフェクト, 報酬
パワー	Power	982	518	エリート, 威力, 争い
報酬	Reward	185	120	功績, 恩恵, 給付
リスク	Risk	190	103	トラブル, 危険, 回避

Igarashi, Okuda, and Sasahara, PsyArXiv, 28 Aug. 2021

J-MFD

- 道徳基盤理論
  - 擁護 (Care)：弱者を守れ
  - 公正 (Fairness)：他人に付け込まれるな
  - 忠誠 (Ingroup)：グループを形成・維持せよ
  - 権威 (Authority)：階層的な社会でうまくやれ
  - 純潔 (Purity)：汚いものから身を守れ
- MFD（道徳基盤辞書）(Graham et al. 2009)
  - J-MFDはその日本語版 (Matsuo et al. 2019, 2021)
  - 例：「侮辱*」→「擁護」「権威」

発表内容

1. テキストマイニングとは
2. 辞書ベースのテキストマイニングの事例

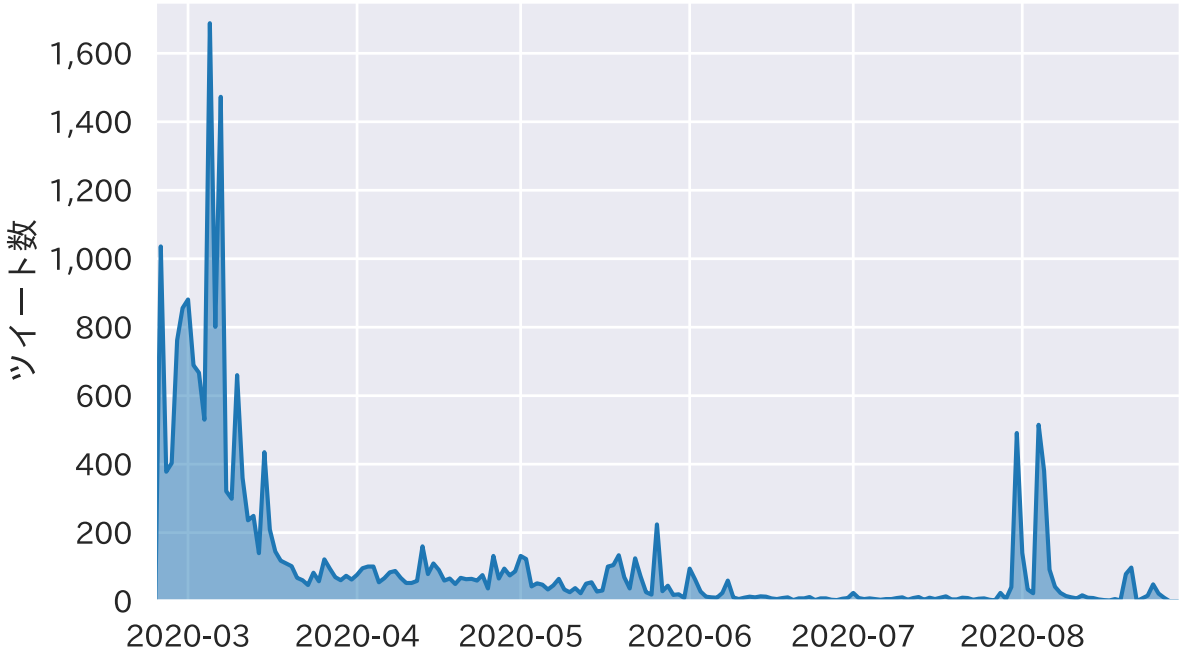
事例 その1

- ツイートからコロナ禍の消費者心理・行動に関する潜在的シグナルを定量化する
  - 転売現象と関連する消費者心理などの変化
  - 転売行動の変化や予兆
- アプローチ
  - 転売に関連するソーシャルメディアのテキストの探索的分析
  - 消費者心理等の定量化に3つの辞書を使用（ML-Ask, J-LIWC, J-MFD）

笹原、奥田、五十嵐, 人工知能学会予行集 2021; Okuda et al. (in prep.)

データ

- 新型コロナに関する単語を含むツイートを収集
- そのうち「転売」が含まれる日本語の投稿を分析
  - 2020/2/25～8/25
  - 22,530ツイート

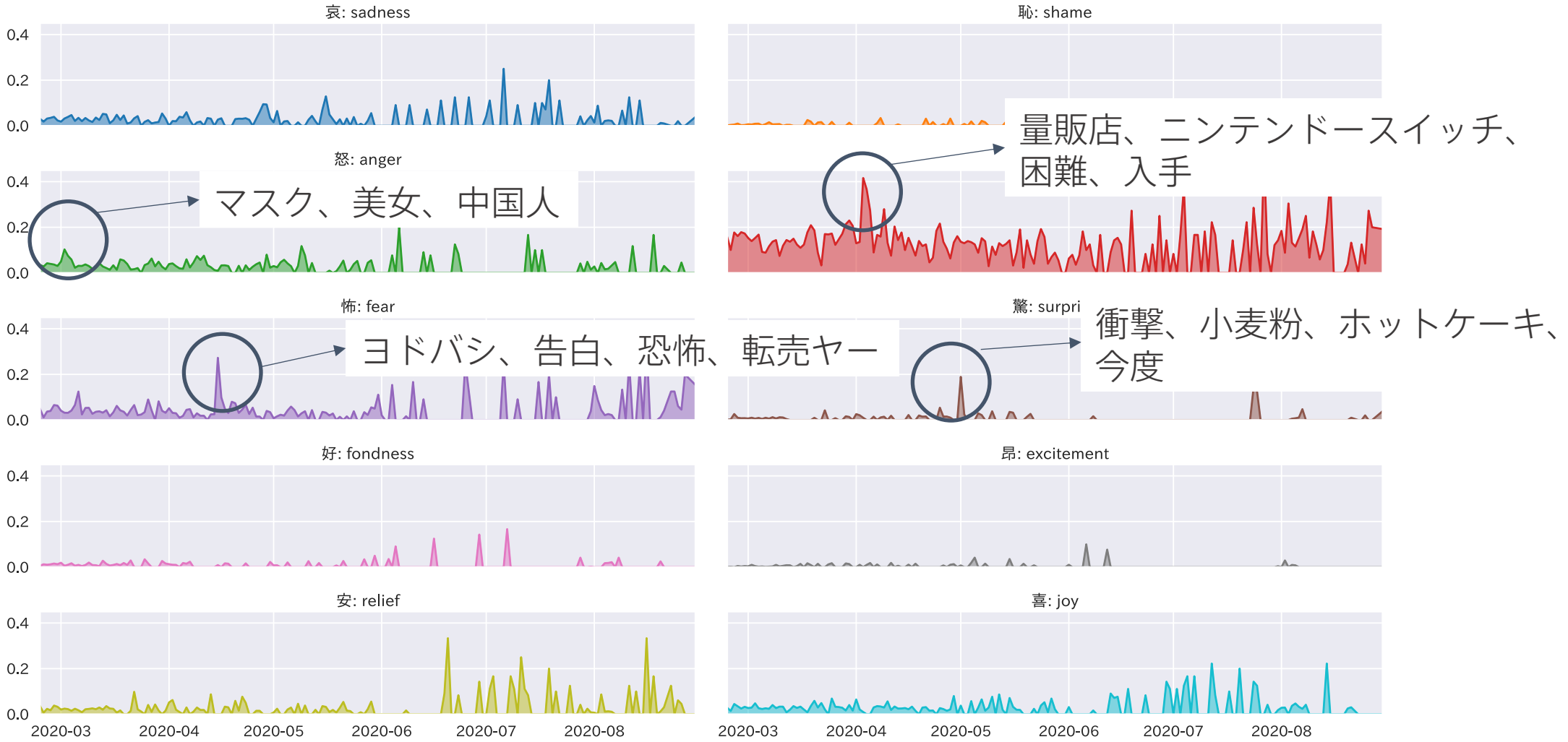


転売と消費者感情

ML-Ask (Ptaszynski et al. 2009, 2017)
により10種類の感情に分類

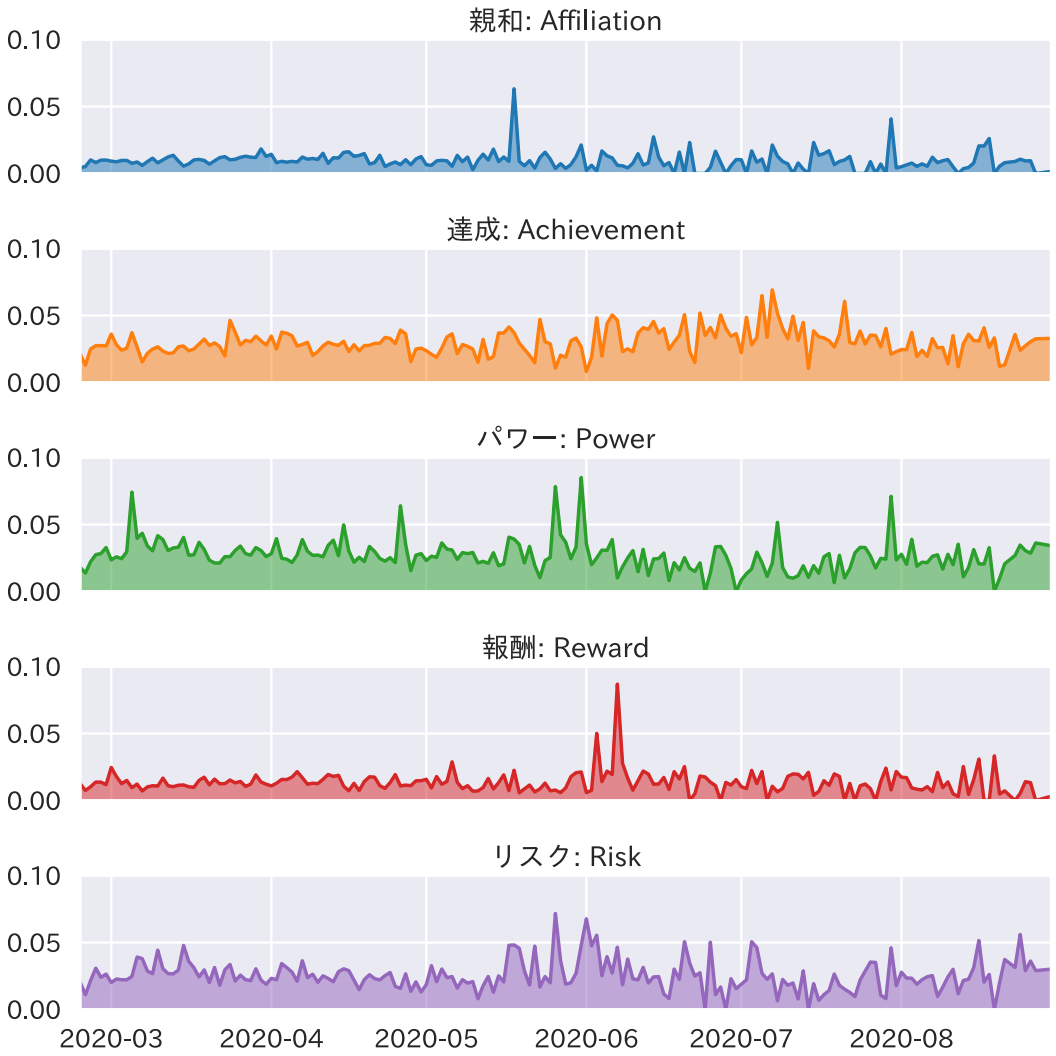
転売と消費者感情

ML-Ask (Ptaszynski et al. 2009, 2017)
により10種類の感情に分類



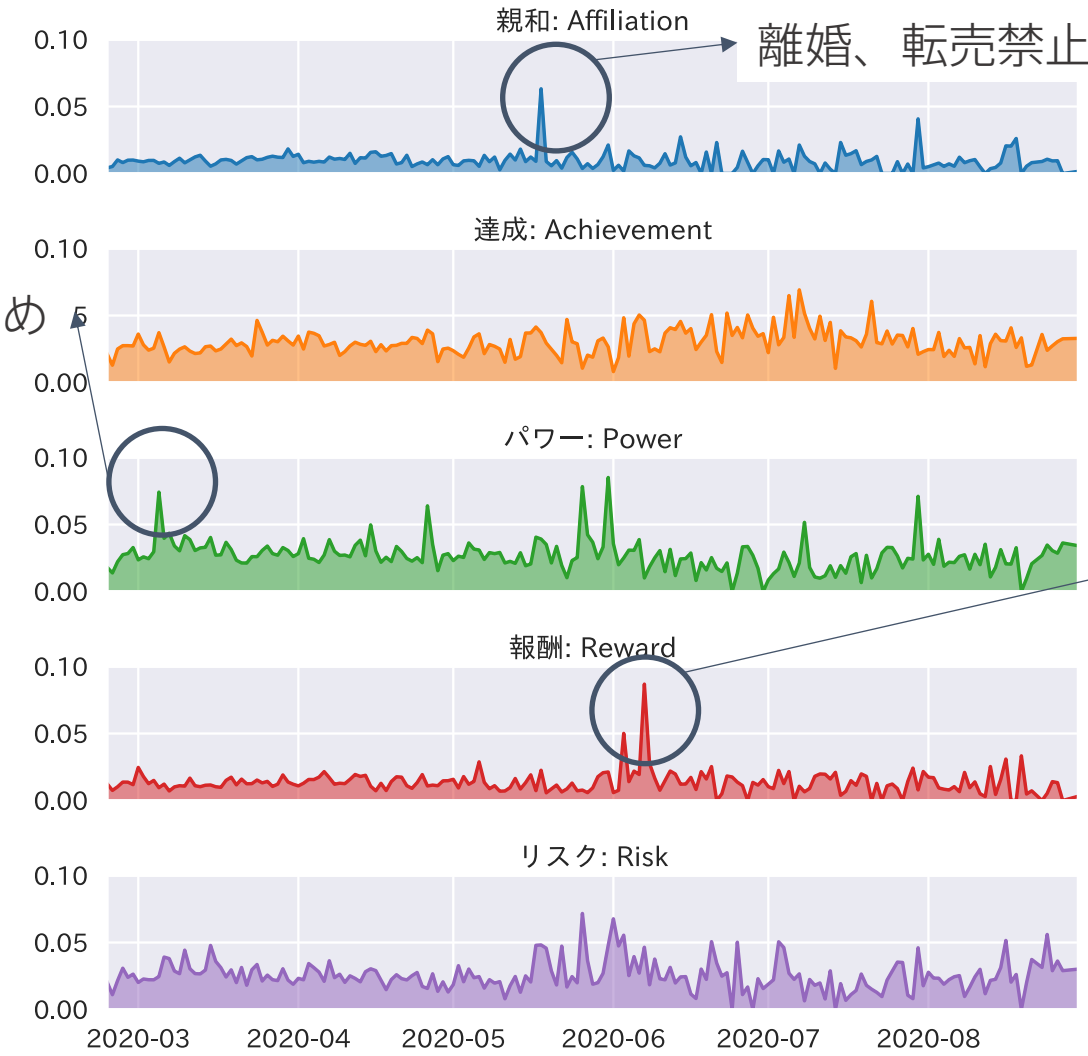
転売と消費者心理

J-LIWC2015の動因に含まれる
5つのサブカテゴリに着目



転売と消費者心理

J-LIWC2015の動因に含まれる
5つのサブカテゴリに着目



離婚、転売禁止、共働き

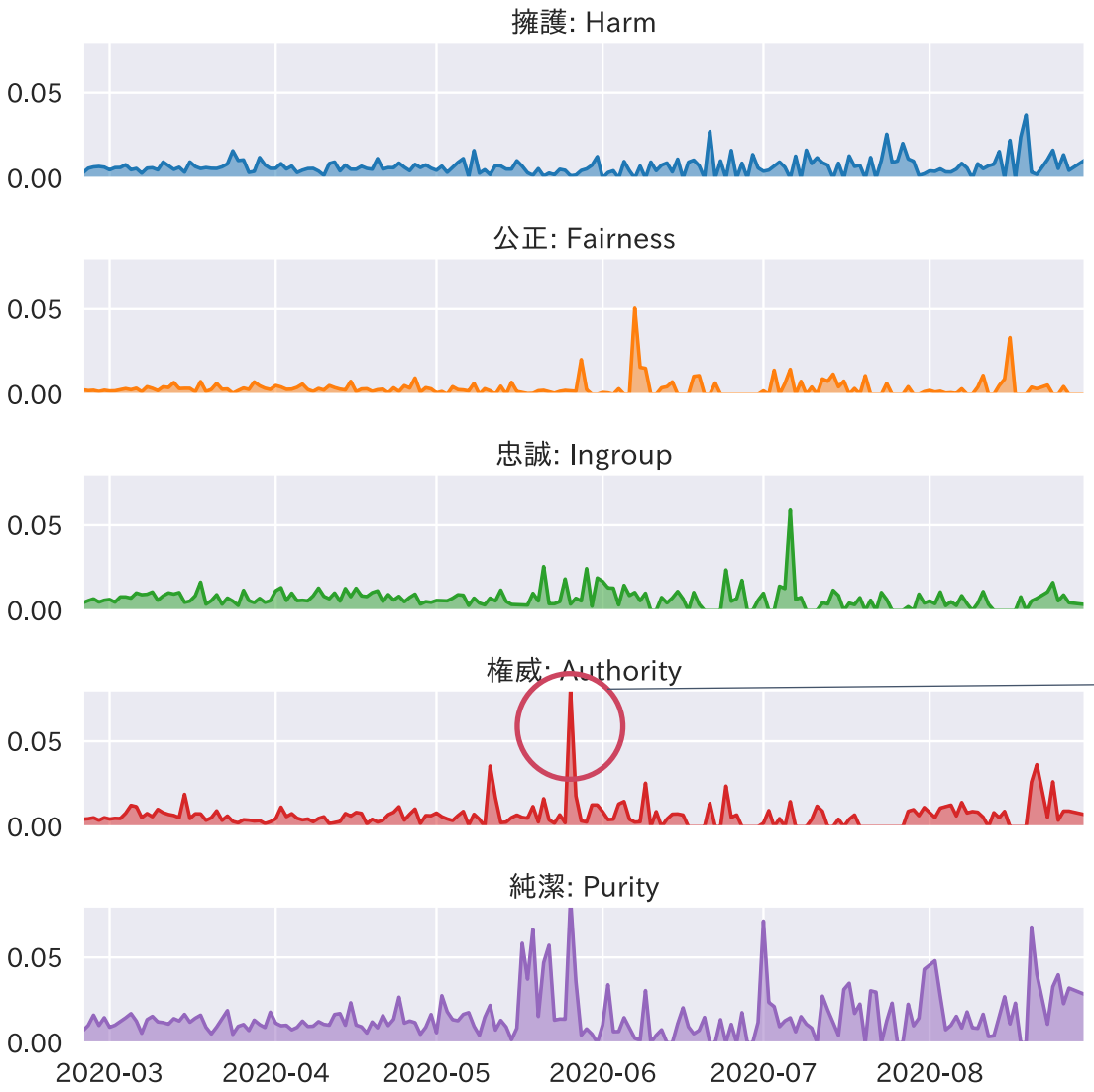
マスク、転売ヤー、買い占め

アングラ、不正受給、
給付金、儲け、ビジネス

転売と消費者の道徳


擁護: Harm
0.05
0.00
公正: Fairness
0.05
0.00
忠誠: Ingroup
0.05
0.00
権威: Authority
0.05
0.00
純潔: Purity
0.05
0.00
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08

転売と消費者の道徳



法律、アルコール、
禁止、消毒液

感染、感染者数、病気

事例 その2

- コロナ禍において反ワクチン運動が深刻化
  - 不安の増幅、ワクチン忌避、陰謀論、転売、詐欺
  - 反ワクチン派の口撃や誘導から身を守り、正しい情報を拡散する
- 反ワクチン運動の口撃を特徴づける

研究手法

- データ
  - 新型コロナに関するツイート（英語860万、日本語200万）のうちワクチン関連語を含む投稿
  - 2020年2月20日～2021年3月31日
- 分析
  - リツイートネットワーク→反ワクチン派などのコミュニティ同定
  - 反ワクチン派の言語的特徴

反ワクチン派と政治的党派性

未発表データのため非表示

Miyazaki et al. (in prep.)

反ワクチン派の返信はネガティブ

未発表データのため非表示

Miyazaki et al. (in prep.)

まとめと今後の課題

- 日英の言語的・文化的比較
  - MFDとJ-MFD
  - LIWCとJ-LIWC
- 辞書ベースの定量化の限界
  - BERTなどのニューラル言語モデルとの比較
- 社会調査と組み合わせる

補足資料

正規化

表現揺れを修正し、ある一定の表記に統一する処理

- 大文字・小文字化：アルファベットは大文字・小文字に統一（例：YouTube→youtube）
- 全角化・半角化：全角・半角に統一（例：ﾘﾝｺﾞ→リンゴ、　２４→24）
- アクセントやウムラウト等の付加記号を削除
- Unicode正規化：結合文字列を解消（例：テ+濁点→「デ」）
- 同義語を統一（例：ファーストフード→ファストフード、国産車→日本車）

基本形への変換

- 語幹処理（Stemming）
  - 活用形、単数形、複数形、派生語を語幹に変換
    例：Family, families, familial → famili
- 見出し語化（Lemmatization）
  - 辞書の見出し語に変換
    例：

「本を読んだ」→「本　を　読ん　だ」→「本　を　読む　だ」

「本を読みました」→「本　を　読み　ました」→「本　を　読む　ます　た」

ストップワードの除去

- ストップワード（不要語）を削除する
  - 機能語：日本語の助詞（「は」や「が」）や助動詞、英語の冠詞や前置詞(a, the)など
  - あまりにも頻度が高い語（場合による）
- 既存のストップワード辞書+αを用意する
  - 様々な言語：RANKS NL Stopwords (https://www.ranks.nl/stopwords)
  - 日本語：Slothlib (https://ja.osdn.net/projects/slothlib/scm/svn/tree/head/CSharp/Version1/)

感情分析

- 辞書ベースの方法
  - VADER：英語の感情値（辞書+ルール）
  - 日本語評価極性辞書（東北大）：単語のネガポジ分類（p, n, e）
  - 単語感情極性対応表（東工大）：単語の感情極性が1～1の数値
  - ML-Ask：10種類の感情（喜、怒、哀、怖、恥、好、厭、昂、安、驚）、ポジネガ分類
- 深層学習ベースの方法（次週）
  - BERTやその拡張モデル