

社会課題の早期抽出に向けた クラウドソーシングによるデータセットの作成とその検討

蕭 喬仁,^{a)} 大知 正直 榊 剛史 森 純一郎 坂田 一郎

概要：社会課題を早期に抽出することは問題が深刻化しその影響範囲が想定外の大きさになることを防ぐ点で重要であるが、これまでその情報源には新聞が多用されてきた。しかし近年、一般市民が社会課題を発見・発信し、それに対して多くの社会の成員が呼応することで社会課題として共有されるという事象が頻出しており、SNS のデータを社会課題抽出の情報源として用いる方がより適切になってきたと考えられる。しかし、SNS 上には社会課題に関連のない投稿も数多く存在するため、これまで SNS を用いた分析は盛んではなかった。そこで本稿では、SNS のデータが社会課題抽出の情報源として使用できるかどうかの検証を目的とし、構築主義に基づく社会課題の再定義を行い、ツイッターからクレーム申し立て活動をクラウドソーシングによって収集することでツイッターのデータが社会課題の抽出に使用できるかどうかの検証を行なった。その結果、ツイッター上にはクレーム申し立て活動にあたるツイートが 2.5% の割合で存在し、そのようなツイートを分析することで社会課題の抽出が行えることを明らかにした。

Building a dataset with crowdsourcing for early social issues extraction

SHIAU CHOUJEN,^{a)} OCHI MASANAO TAKESHI SAKAKI JUNICHIRO MORI ICHIRO SAKATA

1. はじめに

社会課題を早期に抽出することは、問題が深刻化しその影響範囲が想定外の大きさになることを防ぐ点で重要である。これまで、社会課題の抽出のために用いられてきた情報源には、情報の信頼性・多様性・即時性・継続性が高い新聞記事が有用とされてきた [6] が、最近では一般市民が潜在的な社会課題を発見・発信し、それに対して多くの社会の成員が呼応することで社会課題として共有認識されるという事象があちこちで頻出している [8]。そのため、新聞に比べてより即時性が高く、多様な一般市民の声を収集することのできる SNS を社会課題抽出の情報源とすることは有用であると考えられるが、SNS 上には社会課題と関連の無い多種多様な投稿が多く存在し、社会課題に関連する投稿とそうでないものの線引きは曖昧なため積極的な分析が行われてこなかった。

本稿では、SNS のデータが社会課題抽出の情報源として使えるかどうかを検証することを目的に、3.1 で述べる構築主義に基づく社会課題の再定義を行い、クレーム申し立て活動をクラウドソーシングによって収集したのち、得られたデータに対して分散表現に基づくクラスタリングを行うことで、ツイッター上でどのようなクレーム申し立てがなされているか分析した。その結果、ツイッター上にはおよそ 2.5% の割合でクレーム申し立て活動にあたるツイートが存在することが明らかになり、現政権に対するクレーム申し立てが行われているクラスターや会社や学校に対するクレーム申し立てが行われているクラスターが抽出されたことで、クレーム申し立て活動の収集・分析を通じた社会課題の抽出がツイッターのデータで行えることを明らかにした。本稿の貢献は以下のようにまとめることができる。

- 社会課題の再定義を行うことでツイッターのデータを社会課題抽出の情報源として用いることができるようにした。
- 収集したクレーム申し立て活動をクラスタリングすることで、SNS 上でどのようなクレーム申し立てが行わ

¹ 東京大学大学院工学系研究科技術経営戦略学専攻
UT, 〒 113-8656 東京都文京区本郷 7-3-1

^{a)} shiau-choujen293@g.ecc.u-tokyo.ac.jp

れているのかを明らかにした。

- クラウドソーシングによる効率的なデータ収集の方法を提案した。

2. 関連研究

社会課題抽出に関する研究の多くは新聞や政府公表の白書のデータにテキストマイニングの技術を用いる形で行われてきた。新聞や白書に記されている内容の多くは社会課題に関連し、特に新聞は信頼性・多様性・即時性・継続性を兼ね備えた情報媒体であるからだ。橋本らは日経新聞新聞記事から「事件、事故、災害」に関する記事を抜き出し階層的クラスタリングを適用し、クラスタの密度と中心度をもとに重要なクラスタを選択することで社会課題の分析に有用な文書を効率的に選択する手法を提案した [6]。Jeong らも韓国国内の主要新聞紙とニュースメディアのニュース記事に文字数や引用の有無、使用された単語によるフィルタリングを行った上で、トピックモデルを適用し、段落毎の混合トピック比率を求め、社会課題とその内容を説明する文章の抽出を行った [1]。多くの研究が新聞や白書のデータを用いているのに対し、久保田らは 2017 年の 6 月と 7 月のツイートに対して、白書のデータからリストアップした社会課題を教師データとしてツイートから固有表現抽出を行う形で社会課題の抽出を試み、得られた社会課題がツイッター上のコミュニティでどのように議論されているのかの分析を行なっている [5]。

他方で、社会課題の定義方法に関してはこれまで、「専門家が描く社会の目標やスムーズな社会機能を阻害するもの」を社会課題と定義する機能による定義方法があった [3] ほか、「人々が共有する社会的基準と現状の生活との乖離」を社会課題と定義する規範による定義方法 [2] や「価値観の異なる集団間の衝突過程」を社会課題と定義する価値葛藤アプローチによる定義方法があった [7]。しかし、機能による定義方法に対しては専門家にもバイアスが存在し、間違いを犯す可能性がある [4] という問題が指摘されており、規範による定義方法と価値葛藤アプローチによる定義方法に対しては、ある集団の何パーセントが同じ基準を支持すればその基準が共有されていると言えるのかという「数のゲーム」に関する問題やその基準を測定することが可能なのかという問題 [3] が指摘されてきた。対して、構築主義では社会課題を「ある状態が存在するとし、それが問題であると定義する人びとによる活動」とする。加えて、構築主義では問題をめぐる人々の活動は「クレイム申し立て活動」と呼ばれ、クレイム申し立て活動とそれに反応する活動の発生が社会課題である条件 [3] としてきた。

構築主義による定義方法はその他の定義方法と異なり、計算機による社会課題抽出に用いることが可能であるため、本研究でもこの定義方法に基づいた社会課題の定義をし、SNS のデータが社会課題抽出に使用可能かの検証を行う。

3. 手法

本章では、本研究では SNS データを社会課題抽出に使用するために構築主義に基づいた社会課題の定義をし、クレイム申し立て活動を抽出をクラウドソーシングによって行う手法について述べる。

3.1 社会課題の定義

本研究では社会課題を以下のように定義する。

- ある事柄が社会課題であるということは、その事柄を問題であるとする人々のクレイム申し立て活動が存在し、それに対してまた別の人々が様々なアクションをしている状態である。

この定義ではクレイム申し立て活動を行う人の属性や人数、クレイム申し立ての内容のそのものは社会課題であるかの条件ではなく、他者からの反応が条件になっている。クレイム申し立ての内容の真偽を確認することは一般には難しいし、人数を条件に入れると、ごく少数の人が問題であるとする事柄は社会課題とは言えないのかという問題があるからだ。

この定義にもとづけば、SNS 上でクレイム申し立て活動にあたる投稿を収集し分析することで社会課題の抽出が可能になるが、実際にそのような投稿が SNS 上にあるかは定かではないため、次節以降では、代表的な SNS であるツイッターからクラウドソーシングによってクレイム申し立て活動を収集し、その検証を行う。

3.2 クラウドソーシングによるクレイム申し立て活動の収集

クラウドソーシングによってアノテーションをしてもらうツイートデータには 2018 年 12 月 1 日から 2019 年 2 月 28 日から、日ごとに 20 万件をランダムサンプリングした 197 万件のツイートを使用する。しかし、これらのツイートの中にはクレイム申し立て活動に当たらないツイートも数多く存在しているため、そのままアノテーションを行うのは非効率的である。より、効率的なアノテーションのために本研究では以下のスクリーニングによって明らかにクレイム申し立て活動に当たらないツイートの削除を行う。

- (1) 139 文字以上のツイートの削除
- (2) ひらがな、カタカナ、漢字を含まないツイートの削除
- (3) 絵文字を含むツイートの削除
- (4) アカウント名に BOT, Bot, bot が含まれているアカウントからのツイートの削除
- (5) 期間中の投稿数が 100 件以上のアカウントからのツイートの削除
- (6) 投稿されたクライアントが以下のクライアントに該当するツイートの削除
 - twittbot.net, BelugaCampaignSEA, SHOWROOM-

LIVE, Botbird tweets, 今日のツイライフ, Echoes Act2, TwitCasting, Hotpepper, グランブルー フェンタジー

(1) のスクリーニングは文字数が 139 文字以上であるツイートはツイッター公式 API の制約により省略が行われてしまうため、文意が正しく取れない可能性があるから行ったものである。(5) のスクリーニングは期間中の投稿数が 100 件以上であったアカウントの多くが企業の広報アカウントであり、その他のアカウントは現在凍結されているアカウントであったために実施した。また、(6) のスクリーニングはツイート数が多い上位 30 件のクライアントのツイートを目視で確認し、明らかに bot であるクライアントとクレーム申し立て活動に当たらないものを選んだ。

以上で述べたスクリーニングに加えて、プライバシーや後続の処理の観点から以下のような前処理も行った。

- 改行文字に当たる \n と \r\n の削除
- http および https から始まる url を URL に統一
- メンションなどに含まれる @アカウント名を @USER に統一
- & や >, < をそれぞれ, &, >, < に変換

表 1 クラウドソーシングで設定したオプション

オプション	設定した値
タスク毎の設問数	36 問
チェック設問数	9 問
タスク毎の重複質問回数	7 回
タスクの実施上限回数	10 回
アノテータの年齢	18 歳以上
アノテータの性別	誰でも実施可能
アノテータの居住地域	誰でも実施可能

SNS に投稿された文章を読んで、投稿者が **批判や問題提起**をしているかを教えてください。

質問数：36 問 (不適切な回答を防止する質問 9 問を含む)

いただいた結果は日本語を扱う研究に使用いたします。

【注意事項】

- 何かの記事や誰かの発言の引用のみで **投稿者の意見が含まれていない場合は、「批判や問題提起をしていない」と回答して下さい。**
- R-18 内容を含む可能性があるため、18 歳以上限定となります。

【回答例】

「加藤浩次 南青山住民の「幸せ」価値観のズレを指摘「あさましいし、恥ずかしい」 URL」
何かの批判をしていると判断できますが、記事タイトルの引用のみで書き手の意見が含まれていないので「批判や問題提起をしていない」と回答して下さい。

「パワハラが許される時なんてない。豊田真由子議員、埼玉県警が聴取 URL」
批判をしていて、かつ記事タイトルを引用していると判断できますが、前半部分が書き手の意見と判断できるので「批判や問題提起をしている」と回答して下さい。

アノテーションには Yahoo!クラウドソーシング^{*1} を利用した。表 1 にタスク設計の際に指定した重要なオプションを示している。アノテータの年齢が 18 歳以上であるのはツイートの中には R-18 コンテンツがあるためだ。チェック設問とはアノテーションの質を担保するためにタスク掲載者があらかじめラベルを付与した設問のことである。Yahoo!クラウドソーシングではチェック設問に付与されたラベルが事前に用意した正解と一致したアノテータにのみ謝礼ポイントを進呈するオプションがあり、今回はそれを利用することで、不適切なアノテーションの入稿を防止した。チェック質問として用いたツイートの例は表 2 の通りである。また、タスク説明文に「クレーム申し立て活動」という構築主義における専門用語をそのまま記載することはできないため、説明文の書き方を以下のように変えながら小規模なタスクを掲載して実験を行い、(4) の説明文を使用することにした。図 1 には実際のタスク説明画面を示している。

- (1) 短い文章を読んで、何かに批判をしているかどうかを判断するタスクです。
- (2) 短い文章を読んで、批判や問題提起をしているかどうかを判断するタスクです。
- (3) 短い文章を読んで、書き手が何かに問題を感じているかどうかを判断するタスクです。
- (4) SNS に投稿された文章を読んで、投稿者が何かの批判や問題提起をしているかを教えてください。

3.3 結果

3.2 で示した手順でスクリーニングを行うことで約 200 万件あったツイートを 106 万件にまで削減することができた。この 106 万件のツイートから新たに 38,556 件をランダムサンプリングし、図 1 に示すアノテーションのタスクを実施した。その結果を表 3 に示す。過半数を超える 27,206 件のツイートに対しては 7 人のアノテーター全員がクレーム申し立てにはあたらないと判断しており、7 人ともクレーム申し立てにあたりと判断したツイートはわずか 164 件にとどまった。この中から、過半数のアノテータが批判や問題提起をしているとみなしたツイートをクレーム申し立て活動にあたるツイートであるとする、全部で 1974 件の「クレーム申し立て活動」を取得できたことになる。スクリーニング前語のツイート数の増減を考えると、およそ 2.5% の割合でクレーム申し立て活動にあたるツイートが存在していることがわかった。また、クレーム申し立て活動にあたるツイートの文字数を集計したところ、平均文字数は 62 文字で標準偏差は 34.41 であり、最小文字数は 4 文字で最大文字数は 138 文字であった。このことから比較的短文でクレームを申し立てているものから長文のものまで

図 1 アノテーション作業の説明画面

^{*1} <https://crowdsourcing.yahoo.co.jp/>

表 2 チェック質問として用いたツイートと正解ラベルの例

ツイート本文	付与されたラベル
今日で仕事納めだぜい！…もうやること無いぞ (´・ω・｀)	批判や問題提起をしていない
五臓六腑に染み渡った#ゴールデンファイブヌードル URL	批判や問題提起をしていない
どんな想いで自分の子供虐待するんやろな	批判や問題提起をしている
不倫とか淫行とか貧困調査とかのニュースばかりで気持ち悪い たまには「某人気俳優、森を歩くと小鳥たちが集まってくることが発覚」 くらいの爽やかなニュースが見たい	批判や問題提起をしている

表 3 アノテーションの結果

批判や問題提起をしている と判断された数	0	1	2	3	4	5	6	7
ツイートの件数	27206	5811	2287	1278	793	601	416	164

幅広く取得できていることがわかる。

4. 実験

本章では、3で示したクラウドソーシングによって得られた1974件のクレーム申し立て活動にあたるツイートのクラスタリングを行うことで、ツイッター上でどのようなクレーム申し立て活動が行われているかの分析を行う。

4.1 手法

クラスタリングを行うためのツイートの素性は株式会社ホットリンクが公開している日本語大規模SNS+Webコーパスによる分散表現^{*2}を利用して作成した。具体的には、ツイート中の単語を名詞と固有名詞のみにフィルタリングした後に、各単語の200次元の分散表現をTF-IDFを用いて重み付け平均する形で算出した。ただし、学習済み分散表現に学習されていない単語に関しては、全単語の平均ベクトルによる代用を行なっている。なお、形態素解析には形態素解析エンジン Mecab^{*3}およびmecabipadic-neologd辞書^{*4}を用いた。以上の手順で得られた、ツイートの素性をユークリッド距離によるk-means法で10個のクラスターにクラスタリングする。

4.2 結果

表4にk-means法でクラスタリングされた各クラスターの件数と各クラスター内のTF-IDF値が高い上位10単語をまとめた。クラスター毎に異なる内容に対してクレーム申し立てがなされていることが確認できる。例えばクラスター1では仕事や学校における上司や先生に対するクレームがなされていることが推測でき、クラスター5では安倍首相をはじめ政権に対するクレームがなされていることが推測できる。一方で、クラスター4のように様々な問題に対して感情的なクレームがなされているようなクラスターも存在する。クラスター7はツイートの件数が他のクラス

ターよりも多いことを鑑みると未知語が多く含まれるクラスターである可能性が高い。表5にはクラスター1内の実際のツイートを示している。仕事や学校におけるある事柄に対して問題であるとする旨のツイート、すなわちクレーム申し立て活動にあたるツイートが取得できていることが確認できる。

5. おわりに

本稿では、SNSのデータが社会課題抽出の情報源として使えるかどうかを検証することを目的に、構築主義に基づく社会課題の定義を導入し、クレーム申し立て活動にあたるツイートの抽出をクラウドソーシングによって行った。また、得られたデータに対して分散表現に基づくクラスタリングを行うことで、どのようなクレーム申し立てがツイッター上でなされているかの分析をした。その結果、ツイッター上にはおよそ2.5%の割合でクレーム申し立て活動にあたるツイートが存在することが明らかになり、現政権に対するクレーム申し立てが行われているクラスターや会社や学校に対するクレーム申し立てが行われているクラスターが抽出されたことで、クレーム申し立て活動の収集・分析を通じた社会課題の抽出がツイッターのデータで行えることを明らかにした。今後はこのようなクレーム申し立てにあたるツイートのさらなる拡充を目指すと同時に、クレーム申し立てにあたるツイートかどうかの自動判別技術や具体的な申し立て内容の自動抽出技術の開発を目指していく。

参考文献

- [1] Jeong, D., Kim, J., Kim, G.-N., Heo, J.-U., On, B.-W. and Kang, M.: A Proposal of a Keyword Extraction System for Detecting Social Issues, *Journal of Intelligence and Information Systems*, Vol. 19 (online), DOI: 10.13088/jiis.2013.19.3.001 (2013).
- [2] Robert King Merton, R. A. N.: *Contemporary social problems*, New York : Harcourt, Brace, Jovanovich (1971).
- [3] Spector Malcom, K. J. I.: 社会問題の構築—ラベルリング理論をこえて, マルジュ社 (1977).

^{*2} <https://www.hottolink.co.jp/blog/20190304-2>

^{*3} <http://taku910.github.io/mecab/>

^{*4} <https://github.com/neologd/mecab-ipadic-neologd>

表 4 クラスタ毎のツイート件数と TF-IDF 値の上位 10 単語

クラスター	件数	1 位	2 位	3 位	4 位	5 位	6 位	7 位	8 位	9 位	10 位
1	91	仕事	学校	先生	勉強	上司	会社	子供	家族	職場	事務所
2	171	日本	韓国	中国	日本人	世界	事件	ニュース	捏造	北朝鮮	ロシア
3	138	アニメ	ツイート	広告	動画	情報	画像	名前	無断転載	twitter	配信
4	159	ダメ	嫌い	すぎ	バカ	こいつ	ホント	アホ	ほんま	馬鹿	好き
5	180	安倍	国民	辺野古	発言	韓国	安倍政権	野党	問題	安倍晋三	日本
6	191	問題	必要	関係	理解	レベル	認識	場合	選択	解決	判断
7	383	なのか	めろ	1 人	時間	今日	昨日	予定	100	イベント	メンバー
8	203	問題	質問	運営	以上	お願い	案件	電話	許可	契約	会社
9	274	結局	無理	今日	最近	ゴミ	好き	時間	すぎ	以外	残念
10	184	気持ち	人間	空いて	言葉	他人	いじめ	人生	好き	誰か	勝手

表 5 クラスタ 1 内のツイート例

ツイート本文
RT @USER 「定時までいないといけない」という働き方がおかしくて、 「定時で終わるくらいのワークを各人に割り振るのがマネジメント出会って、 上司から言われたワークが終わったら、報告してさっさと帰れば良いという形にしたほうが良いと思うねんな
RT @USER @USER 無言であることがどのような教育的効果があるの聞いてみたいです。 うちの子の中学も謎のルールが多くあります。いずれ謝罪はやりすぎですね。
RT @USER 子供の学校の道徳の授業を何度か参観したが 「その場で一番偉い人（先生）が何を言って欲しがってるか忖度する能力を養う訓練の場」 として非常によく出来てるので、あれは出来れば欠席しろと子供達に伝える方法を考えてる
RT @USER 家族ネットウヨ化問題ツライ URL

- [4] Yasaveyev, I. G.: The construction of social problems by means of mass communication media. (2004).
- [5] 久保田修平, 大知正直, 長濱 憲, 阪井完二, 榊 剛史, 森純一郎, 坂田一郎: 文脈情報を用いたソーシャルメディアからの社会課題抽出に関する研究 (2018).
- [6] 橋本泰一, 村上浩司, 乾 孝司, 内海和夫, 石川正道: 社会課題発見のための文書クラスタリングとクラスタ評価指標, 人工知能学会論文誌, Vol. 24, No. 4, pp. 333-338 (オンライン), DOI: 10.1527/tjsai.24.333 (2009).
- [7] 中川伸俊: 社会問題の社会学一構築主義アプローチの新展開, 世界思想社 (1999).
- [8] 茅 明子, 奥和田久美: 1D08 社会的問題の顕在性と潜在性が研究成果に与える影響 (It; ホットイシュー gt; コミュニティ型イノベーションの可能性と課題 (1), 一般講演), 年次大会講演要旨集, Vol. 30, pp. 118-123 (オンライン), DOI: 10.20801/randi.30.0118(2015).