様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１８　日現在

機関番号：　３２６８６
研究種目：　挑戦的萌芽研究
研究期間：　2016～2017
課題番号：　１６Ｋ１３１８９

研究課題名（和文）新デジタルメディア時代におけるソーシャル・デザインのためのデータ利活用法研究

研究課題名（英文）Studies on data utilization method for social design in new digital media era

研究代表者
和田　伸一郎（WADA, Shinichiro）

立教大学・社会学部・准教授

研究者番号：２０４５４３６６

交付決定額（研究期間全体）：（直接経費）　　1,100,000円

研究成果の概要（和文）：本研究で試みたのは、Twitterデータを大規模に収集し、統計分析ツールや機械学習（Python）を用い分析することによって、いかなる社会集団が存在し、世論への何らかの影響があるのかを示すことである。Twitterは近年、中傷などネガティブな内容が多いとされ問題視されている。私たちはそうした内容の分析に加えポジティブな内容の抽出も行った。データは、2016年4月に起きた熊本大地震をめぐるデータと、2016年7月に行われた東京都知事選をめぐるデータを収集し分析を行った。
その結果、ネガティブな発言をする人々はごく僅かにすぎず、Twitterの設計上、そうした内容が多く感じることが明らかとなった。

研究成果の概要（英文）：In this research, we collected a large amount of Twitter data, analyzed data using statistical analysis tools and machine learning in Python, and clarified what kinds of groups were present and what kind of words were written.
Twitter is a popular SNS in Japan, but in recent years it has been said that there are many negative contents such as slander and it is regarded as a problem. In addition to analyzing negative contents, we also analyzed positive contents. We collected Twitter data on the Kumamoto earthquake that occurred in April 2016 and data on Tokyo gubernatorial election's election held in July 2016 on a large scale and analyzed them.
As a result, we could clarify that because of the design of User Interface of Twitter, negative contents only felt a lot, there is only a few people wrote negative words.

研究分野：　デジタル社会学

キーワード：　テキストマイニング　機械学習　Python　Word2vec　Twitter　ヘイト・スピーチ　選挙　クラスター分析

１．研究開始当初の背景

近年、コンピューティングのカジュアル化によって、蓄積されるデータ量が膨大な量に膨れ上がっている。こうした状況で、そうしたビッグデータを高速度に処理する環境を構築したうえで、機械学習などを用いて知見を引き出す分析が求められつつある。

本研究では、Ubuntu LinuxOS と、Python を用いて分析環境を構築し、その後の社会のあり方に大きな影響を与える「選挙」、とりわけ今回は、都知事選挙をめぐる Twitter データを大規模に収集し、クラスター分析を行うことで、SNS が選挙に関してユーザーの多様なグループにどのような影響を及ぼしているかについて研究を行った。

２．研究の目的

この都知事選は投票率が高く、有権者の関心の高い選挙だったと言える。全体の投票率は 59.73%であり、前回の 2014 年の都知事選の 46.14%に比べ、13.59 ポイント上昇した。では、都知事選に関して、どのくらい「政治的関心」が高かったのか、いわば有権者の生の声を Twitter データからどれだけ抽出できるか、これを行うことを本報告では目的とする。

すでにいくつかの類似した先行研究が存在する[1]。しかし、それらはいずれも、本研究が用いた 2016 年都知事選をめぐる大規模の Twitter データを収集できておらず、また、詳細に分析できているとはいいがたい。研究は、現時点の日本では報告者の知る限りでは見つかっていない。

３．研究の方法

(1)データの収集について

Twitter データは、ユーザーローカル社の協力を得て、7 月 10 日～8 月 7 日の間（ただし使用したデータは、7 月 12 日から 8 月 1 日とした。理由は参院選の投開票日が 7 月 10 日だったため。図１、２参照）のものを収集した。検索ワードは「小池 OR 増田 OR 鳥越 OR 百合子 OR 寛也 OR 俊太郎」（以下、小池 OR データと略記）、「選挙 OR 都知事選 OR 都知事選挙 OR 知事選挙 OR 知事選」（以下、選挙 OR データと略記）とし、これにヒットする Twitter データを全数（合わせて 6,128,753tweet）、収集することができた。

本報告では、これらのコーパス空間の大部分がノイズデータからなると推測し、だとすると、有権者の選挙への関心が見て取れるクラスターは、ごく小さなものだと考えられる。

したがって可能な限り多くのノイズデータを除去できれば、その分、小さなクラスター群を発見できるのではないかと考え、以下の手順でノイズデータの除去を繰り返し行った。

まずは、Twitter の場合とりわけ公式リツイート（以下、RT と略記）は UI 上でボタンをタップ、クリックすれば、簡単に発信できるため、ノイズデータになりやすい。そこで、それぞれのデータについて、RT の割合を出してみたところ、図１、２のように、多くが RT を占めることが分かった。RT が多いものをいくつか確認したところ、これについてはよく知られているように、スキャンダラスな内容が多く、本報告が求めている、選挙あるいは投票への有権者たちの関心とは無関係であるものばかりであった。そこで、選挙 OR データと小池 OR データそれぞれに、RT を除去したデータを作成し、計４つのファイルを分析対象とすることとした。

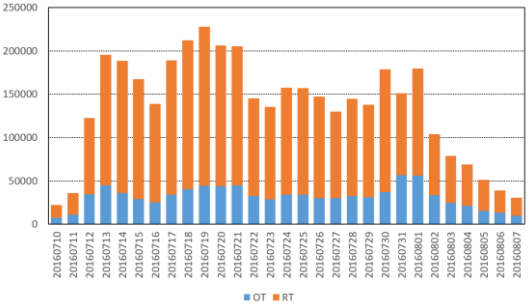

図１．小池 OR データ（OT はオリジナル tweet、RT は Retweet）

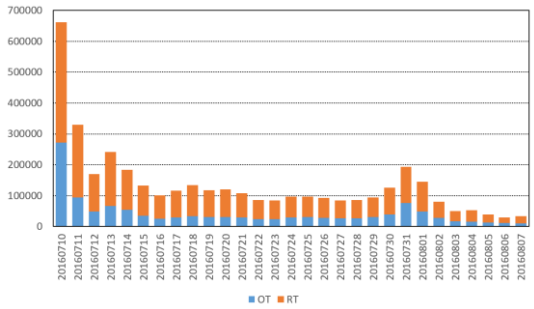

図２．選挙 OR データ

(2) データ・クリーニングについて

次に、tweet に含まれる記事リンク URL や記号など、文字通りのノイズを除去するためにデータ・クリーニングを行った。その後、それぞれのデータに形態素解析、すなわち、品詞ごとに分かち書きを行い、助詞など一語の単語を除去した上で、単語数を算出した（表１参照）。なお、分かち書きには、辞書エンジンである Mecab と、辞書には ipadic 辞書に加えて、Mecab - ipadic - Neologd[2]を用いた。表１にファイル毎に tweet 数と単語数を算出した。小池 OR の RT ありデータ（以下、小池 OR_RT ありデータと略記）から、RT を除去したデータ（以下、小池 OR_RT なしデータと略記。選

[1] 山縣史哉、梅原英一（2018）「平成 28 年度東京都知事選挙の Twitter 分析」、信学技報、電子情報通信学会。木田勇輔（2017）「ソーシャルメディアとポピュリストの動員―2016 年東京都知事選挙における Twitter データの分析から―」椙山女学園大学文化情報学部紀要、第 17 巻。

[2] https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md

挙データの場合も同様とする。）は 17%に減少、選挙 OR_RT ありデータから RT を除去すると、22%に減少した。
表１．

	Tweet数	単語数
小池OR_RTあり	3,517,325	110,850
小池OR_RTなし	614,154	106,041
選挙OR_RTあり	2,611,428	120,535
選挙OR_RTなし	581,156	113,469

このことが意味するのは、RT なしデータでは、出現回数が極端に多い単語が、RT ありデータよりも少ないということである。また RT なしデータで単語数が 4 ファイルともあまり変わらないこと、つまり、それぞれのデータの語彙数にあまり差がないことも分かった。

小池 OR_RT ありデータには 110,850 語の単語が含まれているが、そのうち、最も出現回数が多かった単語は図３にあるように「鳥越」で、1,554,276 回出現している。右端の上位 8 位の単語の数が多くなっており、単語の分布が偏っていることがわかる。こうした傾向は、他のデータすべてに見られた。

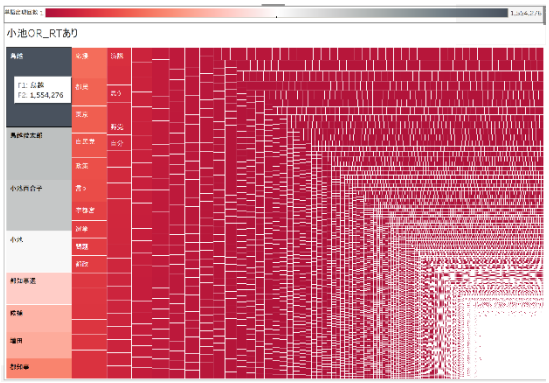

図３．小池 OR_RT ありデータの単語分布

（3）除去したデータの大まかな分類

削除された RT にどのようなものがあったか、おおまかな分類を示しておく。

①－１　ワイドショー的言論の除外。

小池、鳥越といった候補者の名前が入っている tweet、とくに RT では、テレビや週刊誌、ネット記事が取り上げるワイドショー的コンテンツへの tweet、RT が多かった。つまり、政治家に対してというより、有名人に対する言葉であるため、本報告の主旨にそぐわないと判断し、分析対象外とした。

実際、すでに図３で見たように、当選した小池氏よりも、鳥越氏の名前が小池 OR、選挙 OR のいずれのデータでも、最も数が多かった。これは、週刊文春による過去のスキャンダルが取りざたされたことが大きい。

また、小池氏の場合も、石原元知事による「厚化粧」発言、それについての謝罪などワイドショー的なニュースや掲示板でのコメントについての RT、tweet が多かった。

①－２　政治とは相対的に無関係な卑近なトピック

地域の政治とは関連度の低い、これもまたワイドショーでしばしば取り上げられた、オリンピック会場や築地市場移転問題など、日本国民の誰もが口をはさみやすいテーマをめぐる tweet、RT もここでは取り上げなかった。

②　演説を実際に聞いている人の tweet、速報ニュースなどの類へのリアクション。

これは、Twitter という SNS の特性が最も発揮されるコンテンツである。つまり、まさに起きつつある出来事についての tweet（「なう」）、RT である。これには大きく分けて以下の二つの項目が該当した。

1. 速報ニュースや Web 記事が出た直後の反応。
2. 選挙戦で繰り広げられた、東京都内各地での演説をめぐる反応。

１は上の①－２と類似したケースである。２については、データ中に出現する単語の中に、池袋、新宿、八王子など、地名が散見された。これらは各候補の演説会場を示しているが、これらもデータとしてはノイズに近い内容であったため、分析対象としなかった。

③　極右、排外主義者による言論、ヘイト・スピーチの除外

この選挙では、桜井誠氏が立候補し一定数の票を得た。極右の立場をとるこの種の tweet や、掲示板でのコメントへの RT も多くみられたが、この種の偏った言論については、別データとなるが、研究分担者、曹による、熊本地震についての流言飛語についての研究で知見が示されたため、ここでは扱わないこととした。

以上、ノイズデータの除去を行ってきた。次に、都知事選についての有権者の関心が読み取れる tweet、また単語のクラスターの特定を以下の方法を使って探索的に行った。

（4）単語の類似度、クラスターについて

クラスター分析の手法として、Python のライブラリ gensim 版 Word2vec アルゴリズムを用いた。Word2vec アルゴリズムとは「ニューラルネットワークに基づく教師なし学習アルゴリズムであり、単語間の関係を自動的に学習しようとする。word2vec の背景には、同じような意味を持つ単語を同じようなクラスタに配置するという考え方がある。」（ラシュカ（2015=2016））。

①　まずは、4 つのデータを、Word2vec で学習させ、単語をベクトル化した。これによって、単語ベクトル空間の中で、諸単語が互いの間にそれぞれ距離（近さ）をもって埋め込まれることになる。

結果として、互いに関係のある単語だけが残るため、ノイズにあたる単語がさらに除去され、以下の表のように分析対象とする単語数をおよそ半分ほどに絞り込むことができた。

表２．ベクトル化された単語（Point）数

データ	総単語数	Point数
小池OR_RTあり	110,850	76,393
小池OR_RTなし	106,041	52,175
選挙OR_RTあり	120,535	75,872
選挙OR_RTなし	113,469	52,658

このベクトルデータを 4 つのデータすべてについて作成し、それらのデータを用いて、「都」という語を検索したところ、有権者に関わる単語として、いずれも上位に、「(東京)都民」という単語が出現した。ここで注目したのは、この予測変換で、「埼玉都民」、「千葉都民」、「神奈川都民」といった単語が出現したことである。ただし、後者三つの単語についてはそれぞれ単語数が少なかったため、「埼玉県民」、「千葉県民」、「神奈川県民」でデータを抽出し、それぞれの単語数を算出したところ、以下のような結果を得た(カッコ内は、出現回数順のランクを示している)。

表 3．それぞれの単語の出現回数

単語	小池RTあり	小池RTなし	選挙RTあり	選挙RTなし
都民	292956(10)	40254(11)	143722(17)	29106(13)
東京都民	34817(182)	4987(212)	17935(304)	3783(241)
埼玉県民	302(12066)	144(5970)	5654(1195)	375(2740)
千葉県民	289(12325)	112(6996)	464(8079)	239(3932)
神奈川県民	597(8235)	184(5097)	903(5276)	7(34053)
埼玉都民	11(44352)	10(26994)	49(24431)	43(12233)

② 次に、4 つのデータで、上記の単語のうち、この選挙の有権者である「東京都民」を検索語とし近傍語リストを表示させた(「都民」で検索したところ、近傍語にノイズが多く混ざっていたため、こちらは用いなかった)。

なお、近傍語とは、Word2vec によって学習された、ある単語が出てきたときに当該単語が同じく出てくる確率が高い、というところの確率を、近さの数値とするものである。0 に近づくほどその確率が高い、つまり近い＝類似した単語であることを意味する(本研究で作成した、収集した Twitter データを加工し、学習した単語ベクトルデータは、分析精度が一定程度高いものとなりえたが、こうした学習済みデータは「知的財産」に相当しうるという議論があり(角川アスキー総合研究所(2017))、この議論が適切とみなされるなら、作成した学習済みデータを研究成果とみなすことができる)。

興味深い結果として、小池 OR_RT ありデータで「東京都民」に最も近い単語に「良識」という単語が出現した(近さの値は 0.609)。また、選挙 RT ありのデータでも、「良識」は 3 番目に近い 0.672 だった。小池 OR_RT なしデータでは「良識」は 6 番目に近い 0.658 だった(図 4 参照)。ただし、選挙 OR_RT なしデータでは、ランクが低かった。興味深いというのは、後述するように、「良識」を含む tweet からは、東京都民に「良識」をもって候補者を選んでほしい、あるいは、都民自身らが自ら選ぼうという、有権者たちの選挙、投票への関心度を読み取ることができるからである。

続いて、選挙、投票へのとりわけポジティブな関心度を示すクラスターを探索的に見つけるために、次元圧縮アルゴリズムである t-SNE を用いて、学習結果の単語ベクトル空間を三次元に可視化した。

これにあたって用いたのは、Google がオープンソースで提供している、TensorFlow のスタンドアローン版 WebUI、Embedding Projector である。

図 4. 小池 OR_RT なしデータの近傍語

Word2vec で作成した単語ベクトルファイルを、さらに加工した上で、Embedding Projector の WebUI に読み込ませると、設定された語、ここでは「東京都民」という語に近い単語からなるオリジナル・ベクトル空間が PCA(主成分分析)アルゴリズムによって生成される。その後、近い単語が互いに密になりクラスターをなす空間を、t-SNE アルゴリズムを用いて、インタラクティヴにパラメータを変えながら、機械学習させ、最適なクラスター群が可視化できるようにした[3]。

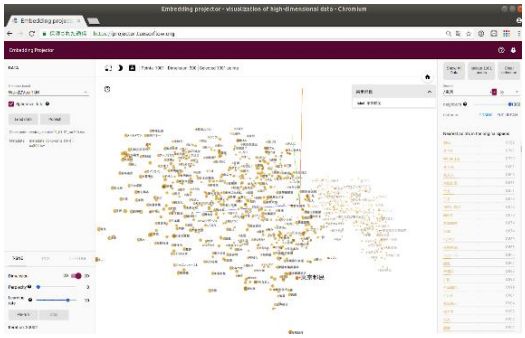

図 5. Embedding Projector で t-SNE アルゴリズムを使ってクラスター分析した 3 次元空間

t-SNE は、「類似するインスタンスを近くに保ち、似ていないインスタンスを遠くに遠ざけながら次元を削減する」(ジェロン(2016＝2018))アルゴリズムであるため、クラスター分析に向いているとされている。

ここで小池 OR_RT なしデータを分析に用いた理由は、ノイズが少なくて、かつ、最も諸単語が多様性をもったデータだったからである。検索語は、「東京都民」とした。

以下では、特徴が出ている三つのクラスターを示しつつ、クラスター内で出現している単語が含まれる tweet のうち、特徴がみられ

[3] OS は Ubuntu Linux18.04、ブラウザは Chromium を使用し、処理速度を上げるために、単語数を 10000points に制限した上で、20002 回学習させた。Perplexity は初期値の 8 のままとした。

たものをそれぞれいくつか挙げる（RT、いいねがあるものはその数を記し、ないものは記載なしとした。また DM とは特定の誰かのアカウントへのダイレクトメッセージを指す）。なおそれぞれの中心語、「良識」、「埼玉県民」、「情弱／情報弱者」という単語を含んだ tweet 数は、それぞれ、497、155、141 であった。これらの tweet は、ほとんどがオリジナル tweet であり、重複した tweet はごく少数だった。

（ｉ）「良識」

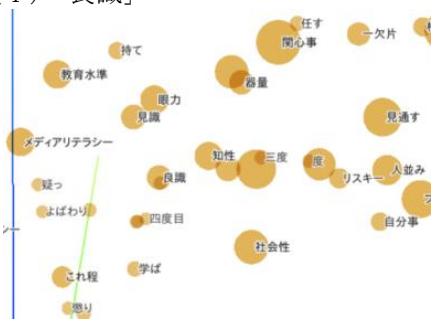

図６．「良識」クラスター

概して、都民に良識に基づいた判断と投票を促す tweet が見られた。

・都民への期待、エールを送るような内容：「健全なる 東京都民の皆様へ[…]2016 年 7 月 31 日全国民が東京都民の皆様の良識を見ています！」

・外野からお願いしたいという上から目線。いずれも、誰かに投票してほしいという含意をもっている。

「都知事選、完璧に野党連合側の思惑通りに事は進んでいます。今のままでは史上最悪の都知事になってしまう。[…]、応援と投票は別と考えるよう、良識ある都民の皆様にはお願いしたい。」RT1

・都民自身：「もうきめました。[…]みなさんも良識をもって判断を！ 都政は国政の場ではなく都民の場！ 自民も野党連合もやり方を間違えている！」DM1

（ii）隣接県民

他道府県民の tweet のほとんどが、「～道府県民なので投票権がない、または、関係ないのだが」という文言を含んでいた。

・「埼玉県民」：「私は隣の埼玉県民ですが、家族が東京都で働いたり学んだりしています。鳥越さんの政策が一番平和的で安心できると感じました。吉井町の、あの美しい田園風景の中で生まれ育った鳥越さん。福岡出身者としても応援しています」RT1 いいね 1

・「選挙権」のある都民をうらやむ内容：「私は埼玉県民として常に誇りを持っているが、今回だけは、小池百合子氏に清き一票を投票できる選挙権が有る東京都民がうらやましい #都民 #都知事選 #東京都知事選 #小池百合子 #都民が決める」RT4 いいね 1

・「千葉都民」：「地元駅で鳥越ビラをまく、という話が入ってきた。東京で働く千葉都民も多いし、東京から来ている人もいる。都外なら鳥越と書いたものを配ってもいいのだ。なるほど」

・「大阪府民」：「長年既得権を貪り続けてきた大阪府市議会議員、大阪府知事と市長。やっと大阪維新の会の登場でその打破の端緒についた。外野の大阪人が言うのも何だが、東京都にもそうなってほしい。東京都民の皆さんの良識ある判断に期待したい。[…]」RT1

・首都圏「有権者」全体への呼びかけ：「民主党は整合性なきマニフェストで政権を取ったが、鳥越俊太郎さんは根拠なき思いつきとクレームのみで、民主党ほどの政策すらない。彼への投票は白紙委任と同じ。民主党政権より政治は後退し、日本のエンジン東京は力を失う。東京で働く神奈川県民、埼玉県民、千葉県民の皆さんも一緒に考えて欲しい。」RT54 いいね 32

・都民に対し疑問を提示するもの：「東京都知事候補の小池さんの公約『満員電車』ってどちらかというと東京都民よりも埼玉県民や神奈川県民等、東京周辺の人たちに対しての恩恵の方が大きくない？そんな投資、東京都民が許すかね」

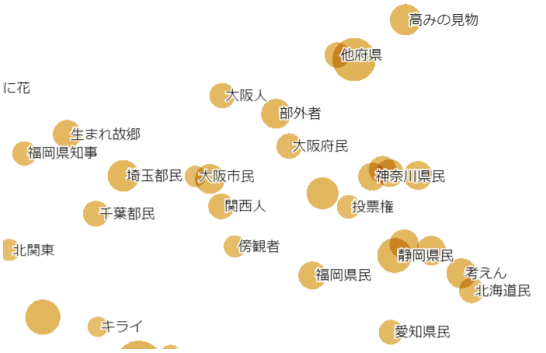

図７．道府県民クラスター

（iii）「情報弱者」、「情弱」

ここで「情報弱者」、「情弱」とは日常的に新聞を読みテレビを観て情報を得ている人々（とくに高齢者たち）のことを指している。

「ネットで情報収集すれば都議会の腐敗が直ぐに分かる。そして増田や鳥越が腐敗都政と非常に親和性が高いのにも気付く。つまり今回の都知事選挙での増田と鳥越の得票数の合計こそは…マスメディアの情報支配下に置かれた情報弱者な有権者の数だろう。」

「鳥越に入る票は、ガチの野党連合か、情報がテレビに偏っている情弱か、といった所で、それが東京でどの位かを観れる選挙だった。で、見えたのが 60 代以上でも鳥越の票の多さ。[…]」

「不都合な情報をカットされたテレビしか見ていない情弱層が選挙で鳥越氏に入れそうでこわいですね。ネットを活用している若い世代に期待したいです」RT2

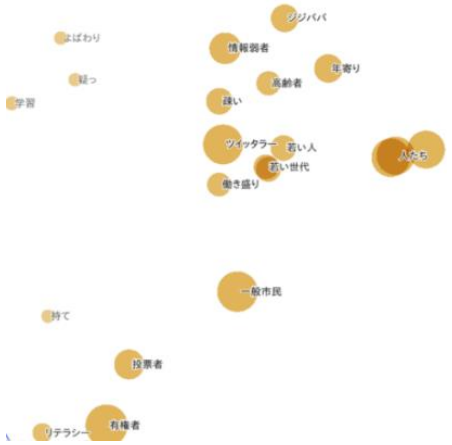

図８．「情報弱者」、「情弱」クラスター

（iv）その他のクラスター

他に識別できたクラスターとしては、先述した保守系のものとして、「桜井誠」／「行動する保守」／「外国人参政権」や、「無党派層」／「取り込み」／「女性票」、また、「期日前投票」／「投票券」／「投票日」／「滑り込み」／「意識調査」、さらに「都民税」／「血税」／「払わ」／「ちょろまかし」などがあった。候補者の公約に関わるものとしては、「原発推進」／「東海村」／「原発マネー」といったものなどがあった。

４．研究成果

以上、有権者の選挙のポジティブな関心が見られるクラスターが複数見つかり、それぞれがどのような文脈をなすものなのか、tweet で確認することができた。

ここに引用した tweet は、すべて OT で、ほとんどが RT されていない。スキャンダラスゆえに数万も RT される tweet や少数のユーザーによって大規模に拡散される保守系の tweet と、ここで扱った RT されず、ほとんどの Twitter ユーザーにすら知られないこうした小市民的な選挙への関心の声との間に極度の非対称性があることが今回の研究で明らかになった。

また、データを分析していく中で、今回収集した Twitter データのコーパス空間が想像以上に大きいものであることが分かった。今回の研究では、具体的に示すことはできなかったが、その空間には複雑な歪みがいくつもあった。しかし、ネガティブな内容が多いことで知られる Twitter にも、ポジティブな内容が、そうした巨大なネガティブ空間の狭間のあちこちに隠れていることを示しえたことで、一定の成果を出せたと考える。

＜引用文献＞

① ラシュカ、セバスチャン（2015=2016）『Python 機械学習プログラミング』株式会社クイープ訳、福島真太朗監訳、インプレス、240
② ジェロン、オーレリアン（2016＝2018）『scikit-learn と TensorFlow による実践機械学習』、下田倫大監訳、長尾高弘訳、O’REILLY、222
③ 角川アスキー総合研究所（2017）『AI 白書 2017』、267

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① 尾鼻崇「ゲームデザインを活用した大学教育の可能性」、『中部大学教育研究』vol.17、中部大学、13-24、2018、（「査読無」）
② 曺慶鎬「インターネット上の災害時「外国人犯罪」の流言に関する研究――熊本地震発生直後の Twitter の計量テキスト分析――」、『応用社会学研究』、60 号、立教大学社会学部、79-89、2018、（「査読無」）

〔学会発表〕（計 5 件）

① 曺慶鎬「災害時における外国／外国人に対する認識の実証的研究 －熊本地震発生直後の Twitter の計量テキスト分析－」、政治コミュニケーション研究会、2018
② Shinichiro Wada ”Artificial Intelligence and Ethical Issues” 、International Joint Workshop、2017
③ 曺慶鎬「インターネットにおける災害時「外国人犯罪の横行」という流言に関する研究 ―熊本地震発生時の TWITTER を事例として－」、多言語社会研究会、2017
④ 尾鼻崇「音響メディアとしてのデジタルゲーム」、日本デジタルゲーム学会 2017 年度夏季研究発表大会、2017
⑤ Shinichiro Wada ”New Trend of Healthy Food Culture Flowing through Social Media in Japanese”、国際共同ワークショップ”Public Spaces in the Digital Age”、2016

６．研究組織

(1)研究代表者
和田　伸一郎(WADA, Shinichiro)
立教大学・社会学部・准教授
研究者番号：２０４５４３６６
(2)研究分担者
尾鼻　崇(OBANA, Takashi)
中部大学・人文学部・講師
研究者番号：００５１６８３３
曺　慶鎬(CHO, Kyongho)
立教大学・社会学部・教育研究コーディネーター
研究者番号：２０７６２８９２