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

37-166828 久保田修平
指導教員 森純一郎 准教授

1 序論

1.1 研究の背景と目的

近年，エビデンスに基づく政策形成 (Evidence-Based Policy Making, 以下 EBPM) の重要性が指摘されている．EBPM とは科学的な手法による客観的根拠（エビデンス）に基づいて，政策の企画立案やその評価及び政策への反映などを行なって行くべきだ，という考え方である．少子高齢化や人口減少など未来の不確実性が高まっている現代において，エビデンスに基づいて社会的な課題を素早く把握し，限られた資源の中で効果的な施策の選択と実行をしていく重要性が増しているのだ．

EBPM に関する研究はまだ数としては少ない [4] が，その数少ない研究の 1 つに Ito らの研究 [1] がある．Ito らは，経済的なインセンティブを与える場合とモラルへの働きかけのそれぞれが電力消費の減少にどれほどの効果を持つのかの検証を行い，金銭的なインセンティブを与えた方がモラルに訴えた場合に比べ消費電力が減少することを明らかにし，金銭的なインセンティブを与える政策の方がより効果的であることを示した．

他方，近年ではソーシャルメディアの発達に伴い，人々は自分の意見をオンライン上に発信するようになった．人々が社会的に話題になっているニュースや個々の抱える課題などもソーシャルメディア上に発信することによって，ソーシャルメディア上には個々人の日々の出来事や関心事などに関するデータが大量に蓄積されることになった．ソーシャルメディア上の情報には人々の興味や関心が多く埋もれており，そうした情報をうまく分析することで EBPM や世論調査など政策分野へのエビデンスとしての活用への可能性が十分秘められている．

一般に，自然言語処理の世界で特定のカテゴリの語 (組織名や国名など) を文中から抽出するタスクは固有表現抽出と言われる．すなわち, 前述の議論を踏まえれば，ソーシャルメディア上の社会課題に関する固有表現抽出を行うことができれば政策分野に対するエビデンスとして期待ができる．ところが，ソーシャルメディア上の投稿に対して従来の固有表現抽出のモデルを適用するには 2 つの課題が存在する．1 つ目はソーシャルメディア上のコンテンツのノイジーさに起因するものだ．ソーシャルメディア上のコンテンツは短文で文法なども少し崩れたノイジーなものであることも多く，従来の手法をただ適用するだけでは十分とは言えない．2 つ目の課題として，ソーシャルメディア上で投稿を行なったユーザの信頼性や他ユーザとの関係性といった社会的背景を考慮に入れられていないことだ．特に社会課題においては投稿を行なったユーザの信頼性などの社会的背景を考慮に入れた上で抽出することが重要であるが，そうした社会的背景を加味した上での固有表現抽出を行なった研究はあまりない．

そこで，本研究ではソーシャルメディアにおける社会課題に特化した固有表現抽出モデルを作成することを目的とする．そのために前述した 2 つの課題を解決するための新しい特徴量を設計し利用する．具体的にはコンテンツのノイジーさについては，係り受け解析を用いることでより直接的な文脈を把握し，社会課題抽出の精度をあげる．また，ユーザの社会的背景を入れた分析についてはソーシャルメディアからユーザのネットワーク上の特徴量をいくつか定義し解決を試みる．

また，ソーシャルメディア上の情報を用いることで抽出された社会課題がどういった特徴を持つユーザ群に関心を持たれているかを抽出することも可能だ．そうした情報を用いることによって抽出された社会課題に対して社会的な重要度であったり，関心の広がりであったりを評価することが可能になる．本研究では，そうした抽出された社会課題がどういった人に関心を持たれているかといった情報も含めて整理することで社会課題の全体図を概観することを最終的な目的としたい

本研究では，投稿量の多さととネットワーク情報抽出できるということからソーシャルメディアとして Twitter のデータを分析する．

本研究の研究の目的は以下に要約される．

- ソーシャルメディアからの社会課題抽出モデルの作成
- ユーザクラスタリングと社会課題の関係分析

2 手法

本研究は大きく 2 つの要素で構成されている．1 つは，社会課題を抽出するモデルの作成で，もう 1 つは社会課題に関心を持つユーザの層を特定するためのユーザのクラスタリング分析である．

2.1 社会課題抽出モデル

元来，文書中から人名や地名といった固有表現など特定の表現を抽出するタスクは，固有表現抽出というタスクとして自然言語処理の中で研究がなされてきた．ゆえに，本研究は社会課題固有の固有表現抽出器の構成に関する研究だと言える．固有表現抽出は，各単語に固有表現タグをラベリングして行くタスクとして考えることができる．固有表現タグとして，本研究ではIOB タグを使用する．IOB タグとは，ある単語が固有表現であるかを表現するタグであり，固有表現の先頭の単語を B タグ，同一の固有表現であり，B タグに連続して繋がっている単語に対しては I タグ，そのほかの単語に対しては O タグをふっていく．このように各単語を IOB の固有表現タグに分類していくことによって固有表現を文章中から抽出することが可能となる．

2.1.1 使用する特徴量

まず，基本的な特徴量として「前後 4 単語」「前 4 単語の固有表現タグ」「前後 4 単語の品詞」「前後 4 単語の文字種」を利用する．図 1 にその概略図を示す．ここでは,「ゴミ問題」という社会課題の最初の単語である「ゴミ」にタグづけを行う場合の図を表している．ここでいう「前後 4 単語」というのは単語の word2vec によってベクトル化した特徴量を意味する．分散表現を利用することで意味的な情報を加味することができ，Twitter という文法も少し崩れたノイジーなものであること，社会課題という特有の文脈で現れてくる表現を抽出することを加味すると，本研究の精度向上に寄与すると期待される．また，前述の通り Twitter のネットワーク情報やユーザ情報などもモデルの特徴量として組み込むことにする．具体的にどのような特徴量を構成していくかを以下に述べていく．

位置	i - 2	i - 1	i	i + 1	i + 2
単語	都会	の	ゴミ	問題	は
品詞	名詞	助詞	名詞	名詞	助詞
文字種	その他	ひらがな	カタカナ	その他	ひらがな
タグ	O	O	B	I	O

図 1: 基本モデル

2.1.2 単語の分散表現

本研究では Mikolov らの skip-gram Negative Sampling(SGNS) モデル [2] を用いて単語の分散表現を構成した．そして，この SGNS によって作成された単語の分散表現を社会課題抽出モデルの特徴量として採用している．

まずは SGNS の基礎となる skip-gram モデルについて説明する．skip-gram モデルは，英語では“words that occur in the same contexts tend to have similar meanings“ と表現される分布仮説という考えに基づいて構成されたモデルである．つまり，周りの文脈を利用することで中心にある単語の意味的な表現を獲得しようとする考え方である．ある語からその周辺の単語を予測する学習を繰り返すことによってその単語の意味的な表現を獲得するモデルだ．SGNS は skip-gram モデルにおける目的関数に近似式を利用し学習を高速化させている手法だ．

2.1.3 文脈表現

社会課題が現れる文脈をノイジーなデータの上で学習する際，社会課題のすぐ近くにある単語を利用するよりも，より直接的な形で社会課題の置かれている文脈を特徴量として加えたほうが精度が上がるのではないかと期待できる．社会課題の係り受け情報を利用することで，その社会課題のツイートにおける文脈的な位置付けをよりピュアな形で抽出することができ，それによって社会課題抽出の精度向上につながると考えられるからだ．係り受け情報とは，文章中における文節間の「係る/係られる」の関係性のことで，自然言語処理の世界において係り受け解析というタスクで研究がなされてきた．よって，本研究では，上述の特徴量に加えて，単語の係り受け情報を特徴量として用いることにする．より具体的には，係り受け解析を行うことで，注目している単語が「係られている文節」と「係っている文節」を抽出し，双方の分散表現を特徴量として加えることにする．文節はたいてい複数の単語で構成される．ここでは文節の分散表現として文節を構成する単語の分散表現の平均値を用いる．

2.1.4 ネットワーク特徴量

Twitter のデータからユーザが Twitter 上で引用しているメディアや普段会話しているユーザなどの情報を抽出することも可能だ．本稿で目的とする社会課題抽出のタスクにおいて，これらの情報が社会課題であるかを判定することに役立つ可能性は高い．ツイートを行なったユーザが普段会話しているユーザや普段引用しているメディアなどの情報から，そのユーザの社会課題に対する関心度などの情報を得られると考えられるからだ．

Twitter 上においてユーザと他のユーザやメディアとのインタラクションは 5 つに分類することができる．「ユーザにメンションする」「ユーザからメンションされる」「ユーザを RT する」「ユーザから RT される」「メディアを引用する」の 5 つである．さらに，本研究ではユーザ及びメディアとのインタラクションに関して，大きさと幅によって整理することにする．例えば,「ユーザからメンションされる」というインタラクションにおいては，どれだけ多くのユーザからメンションされたかが幅で，ユーザの重複を許してどれだけ多くメンションされたかが大きさであると定義する．社会課題に対する関心度を表現する上でその人の考えの多様性や社会性などの指標が有効であると考え，どれだけ多くの人の考えを取り入れているかを表現する上で幅という考えを導入した．一方，大きさに関しては，どれだけインタラクションを行なっているかはその人の社会との関わりの量を表現していると考えられ，これもまた社会課題に対する有効な指標になるのではないかと考え，このような分類を行なった．

以上の着想を踏まえて，本研究で利用する特徴量を以下に示す．

表 1: ネットワーク特徴量

使用するネットワーク特徴量
RT したユーザの数
RT した回数
RT されたユーザの数
RT された回数
メンションしたユーザの数
メンションした回数
メンションされたユーザの数
メンションされた回数
各メディアの引用回数 (TOP 200)

2.2 予測モデル

本節では，上で述べてきた特徴量を再度整理し，社会課題抽出モデルの構成について記述する．本研究では，各特徴量の効果を測定するため，3 つの実験を構成する．表 2 にその 3 つの実験を示す．実験 1 は従来の研究と同様に「品詞・文字種・単語分散表現」を利用し，実験 2 ではそれに加えて「係り受け情報」を特徴量にする．さらに，実験 3 では実験 2 に加えてネットワーク特徴量まで加えて特徴量とする．

表 2: 実験の構成

実験	特徴量
実験 1	品詞・文字種・単語分散表現
実験 2	実験 1 ＋ 係り受け情報
実験 3	実験 2 ＋ ネットワーク情報

本研究では，固有表現タグの予測モデルとしてロジスティック回帰を使用することにする．ロジスティック回帰とは以下の式で表される確率モデルに従って，ラベル推定を行うモデルである．ここで w_i はパラメータを表し，x_i は特徴量を表している．

$$p_i = \frac{1}{1 + exp(-\sum_i w_i * x_i)}$$

2.3 ネットワーククラスタリング

ユーザをクラスタリングするためにメディアとユーザ間の会話ネットワークを利用する．また本研究ではユーザが引用する web 上の情報ソースも含めてクラスタリングを行うことでより興味関心のあるユーザ群を抽出できると同時に，クラスタを特徴付ける上でも有効であると考え，メディアとユーザの混合のネットワークを構成し，クラスタリングする．具体的には図 2 に示されているように，web 上のメディアをあるユーザが引用している場合に，そのメディアとユーザ間にエッジを貼り，またユーザ間にメンションがあった場合にユーザ間にエッジを貼ることでネットワークを構成する．

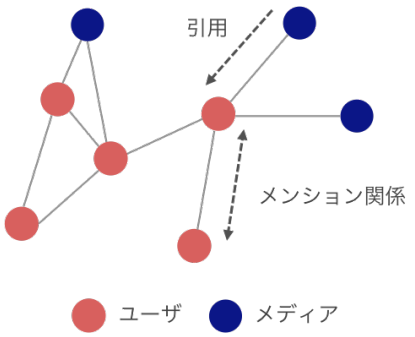

図 2: ネットワークの構成方法

このように構成したネットワークを Louvain 法 [3] を用いてクラスタリングすることで，同様の興味を持つユーザコミュニティを抽出することができる．Louvain 法は分割の精度を表す Modularity を高速に高いスコアで分割する手法として知られている．Modularity とはコミュニティ内のリンクが密で，コミュニティ間のリンクが疎であるほど高いスコアとなる指標である．あるネットワークに関して，頂点集合を V，V のある分

割を $D = \{V_1, V_2, .., V_k\}$ とするとき，この分割 D に対する Modularity は以下の式で表される．

$$Q(D) = \sum_{i \in C} \left(\frac{e_{ii}}{2m} - \left(\frac{\sum_{j \in C} e_{ij}}{2m} \right)^2 \right)$$

ここで C はコミュニティの集合であり，e_ij はコミュニティ i からコミュニティ j に貼られているリンクの数，m はネットワーク全体の総エッジ数を表している．

2.3.1 クラスターへの特徴づけ

各クラスターを特徴付けるため，各クラスターに所属するユーザの Twitter 上のプロフィール欄で使用されている言葉を抽出し，TF-IDF 値で各単語の重要度を算出した上で，スコアの高い言葉をもってそのコミュニティを特徴付けることにする．具体的には，同一クラスターに所属するユーザーのプロフィール文を全て結合して 1 つの文書とみなし，各クラスターのプロフィール文に含まれる単語の TF-IDF を算出している．その TF-IDF 値が高い単語がそのクラスターを特徴づけている単語として抽出されることになる．

2.3.2 社会課題の分類

さらに本研究では，ユーザクラスタの情報を用いて社会課題を分類することを試みる．図 3 が本研究における分類のフレームワークである．ツイート数の量と関心のあるコミュニティの偏りによって，社会課題を 4 つのタイプに分類する．広く色々なコミュニティにツイートされていて，ツイート数が多いものを「1. 大きな社会課題」，逆にツイート数が少ないものを「2. 社会課題の芽」，また，特定のコミュニティの偏ってツイートされているもので，ツイート数の多いものを「3. 一部の人にとって重要な社会課題」，逆にツイート数が少ないものを「4. 一部の人にとっての社会課題の芽」というように名付けた．

	全体に広く分布	特定のコミュニティに偏っている
ツイート数が多い	①大きな社会課題	②一部の人にとって重要な社会課題
ツイート数が少ない	②社会課題の芽	④一部の人にとっての社会課題の芽

図 3: 社会課題の 4 分類

本研究では，社会課題を定量的に 4 つに分類できるように，各軸を定量的に評価する．まず，各社会課題におけるコミュニティの偏っている度合いの指標としてはエントロピーを採用する．エントロピーとは以下の式で表される指標であり，コミュニティに偏りがないほど値が大きくなる性質がある．

$$entorpy = -\sum p \log p$$

さらに，2 次元上で社会課題をマッピングするために，ツイート量に関しては単純なツイート数を log をとったものをさらに全体で標準化したスコアを採用し，コミュニティの偏り度に関してはエントロピーを全体で標準化したスコアを採用する．

3 実験と結果

3.1 データセット

本研究では，2 種類のデータを利用する．社会課題のデータと社会課題を抽出する Twitter データである．本研究では正解となる社会課題単語として，2017 年 6 月に閣議決定された「未来投資戦略 2017」に記載されている社会課題に関する単語を学習に利用するための社会課題として抽出した．また Twitter データに関しては，全ユーザから 10%をサンプリングし，そのユーザによって 2017 年 6 月から 2017 年 7 月の間に投稿されたデータを利用した．対象となるツイートの総ユーザ数は 13862496 ユーザで，総ツイート数は 264114133(6 月：123380307 ツイート， 7 月：140733826 ツイート) ツイートであった．

3.2 社会課題抽出モデル

3.2.1 モデルの評価

モデルを評価する上で，選定した社会課題単語のうち 1 つの単語をテストデータに利用し，残りを学習データに利用している．具体的には，学習データに利用する社会課題が含まれているツイートと，社会課題が含まれていないツイートを 1:1 で混ぜたツイートデータを学習データとして学習し，テストデータに利用する社会課題が含まれているツイート全てを精度評価のためのデータとして利用している．テストデータで使用する社会課題として,「技術開発」「安全性」「競争力」「規制改革」「見える化」「サイバーセキュリティ」「労働生産性」を対象とする．学習データとして社会課題を含むツイートを 30000 ツイート，そうでないものを 30000 ツイートを利用している．テストデータに関しては対象とする社会課題を含むツイートを全てテストデータとして利用した.「技術開発」「安全性」「競争力」「規制改革」「見える化」「サイバーセキュリティ」「労働生産性」

を含むツイートは，それぞれ 131,419,805,5389,3958,50 ツイートであった．

3.2.2　社会課題抽出モデルの結果

表 3 に本実験の結果を示す．その結果，実験 1 と比べ実験 2 と実験 3 の精度向上が見られた．特に precision で大きく数値をあげたことがわかる．また実験 2 と実験 3 では精度においてそれほどさが見られなかった．

表 3: 実験結果

実験	precision	recall	f 値
実験 1	0.192	0.149	0.166
実験 2	0.529	0.120	0.19
実験 3	0.532	0.123	0.193

3.3　ネットワーククラスタリング

本研究では Louvain 法を用いて，21 のユーザクラスターに分割した．そして分割されたクラスターを用いてツイート数と関心を持っているクラスターの偏り度で社会課題を分類した．ツイート量とクラスターの偏り度を 2 次元にマッピングした図が図 4 である．横軸がエントロピーであり，縦軸がツイート数を表している．右上の象限が「1. 大きな社会課題」，右下の象限が「2. 社会課題の芽」，左上の象限が「3. 一部の人にとって重要な社会課題」，左下の象限が「4. 一部の人にとっての社会課題の芽」に対応している．

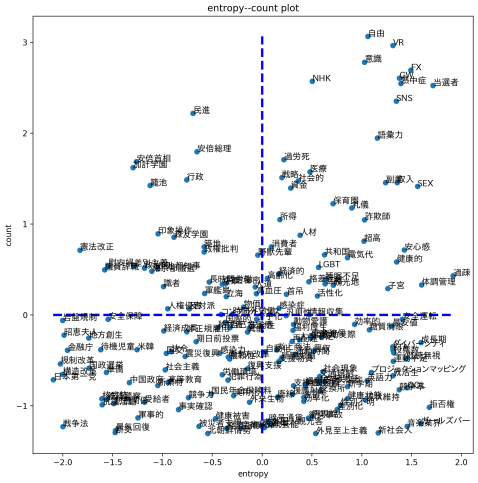

図 4: 社会課題マッピング

4　考察

4.1　抽出モデルにおける係り受け情報の効果

F 値を比較すると，係り受け情報を利用したモデルの方が精度が良いことがわかる．特に，precision の精度改善が大きい．precision は社会課題だと予測したものの正答率のことであり，これは係り受け情報を利用することによって社会課題が登場する特定の文脈をモデルによって学習できたことを示している．Twitter のようなノイジーなメディアでは文法や単語が崩れていることが多いため，係り受け情報を利用することによって，そうした影響を排除し，直接的に社会課題がどういった文脈にあるかを捉えられた結果であると解釈することができる．

4.2　抽出モデルにおけるネットワーク特徴量の効果

実験 2 と実験 3 の F 値を比較すると，実験 1 と実験 2 ほど F 値は離れておらず，あまり精度向上は見られていない．ネットワークに関する特徴量が社会課題推定にあまり寄与していなかった可能性を示唆している．

はじめに，ユーザとのつながりに関する指標について考察する．表 4 はユーザとの繋がりの指標の重みを表している．重みとしては RT された/した回数が唯一正の値を持っていることがわかる．他のユーザを引用したり，他のユーザに引用されることが多いユーザは社会課題に関心がありそうだと言えるかもしれないが，重みとしてはかなり低いので，信頼性には欠ける．ほとんどの指標は効果がないまたは，マイナスの効果を持っていると言っていいかもしれない．

こうした結果になったのは，Twitter のユーザに 1 つの原因があるかもしれない．Twitter には不特定多数のユーザに対して，メンションを送ったり，RT したりする企業や個人またはスパムアカウントのようなものまで存在している．本研究では，ネットワーク特徴量からそのユーザの社会的背景を表現できるという仮説のもとこの特徴量を利用したが，前述のことが背景としてあり，今回の指標では正当にユーザの信頼性といった社会的背景を学習することができなかったことが，原因としてあるかもしれない．ユーザのインタラクションが自然なものなのか不自然なものであるのかと言った分析の上でこうした指標を構成し，分析することでより正しい重みを見いだせる可能性がある．

表 4: 実験結果

ネットワーク特徴量	重み
RT された回数	1.43e-3
RT した回数	5.11e-4
メンションされた回数	2.89e-4
メンションした回数	-1.03e-3
RT したユーザの数	-2.02e-3
メンションされたユーザの数	-2.02e-3
RT されたユーザの数	-1.03e-2
メンションしたユーザの数	-2.76e-2

4.3 ネットワーククラスタリングに関する考察

ネットワーククラスタリングの分析では，ユーザの全体を外観し高校生や大学生といった若者が多いことや音楽・アイドル・ゲーム・アニメといった趣味でつながっているクラスターが多くいたり，かたや政治的主張の強いクラスターも多くいることなどを確認した．これは本研究を政策分野へ応用する際にとても重要だと考えられる．今回抽出した社会課題はあくまで Twitter の中におけるものであるからだ．決して全国民の中でどういった人々が興味を持っているかということは言っていない．だからと言って，この分析に意味がないわけではない．10 代 20 代の若者の間では約 6 割の人々が Twitter を使用しており，普通の世論調査では扱えないほど多くの人々の関心を分析の対象にできることは大きな魅力だ．高齢の人であっても多少興味は偏っている可能性はあるが多くのユーザがいる．重要なのは，どういったユーザがどういった関心を持っているかということを把握することである．そういった意味で今回の分析によって考えの偏りや興味の偏りも含めて，どういったユーザがどういった社会課題に関心を持っているかを明らかにできたことは意味のあることであった．

5 結論

本研究では，Twitter からの社会課題抽出とユーザクラスターと社会課題の関係分析を行った．その中で，Twitter というノイジーなメディアにおいては係り受け情報を利用することが有用であることが示された．また，本研究で使用したネットワーク特徴量では社会課題推定にはあまり寄与しないことがわかった．また，ユーザクラスタとの関係で社会課題を定量的に評価することによって，社会課題の全体図を概観する方法を提示した．このマッピングから人々の抱く社会課題の全体像を把握することが可能になり，政策のリサーチなどに活かせることが示唆される．

最後に，本研究の課題と今後の研究の方向性について述べたい．社会課題抽出モデルでは実験結果とその考察でも述べたようにネットワークから導出した特徴量はあまり効果がないことが結果として得られた．考察において，ユーザとのインタラクションに関する指標はユーザの質が正しく精査されておらず，正当な指標となっていないことが原因であり，メディアに関してはあまり社会課題推定に効果がないことが原因だと論じた．仮にこのことが正しいのだと仮定すれば，ユーザのインタラクションに関する指標に改善の余地が存在している．改善案としては 2 つあげられる．1 つはユーザの中で平均から大きく離れてユーザとインタラクション (メンションを送るなど) をするユーザを外して考えるということである．このようにすることで自然な社会ネットワークが構成できると考えられる．2 つ目としては，ユーザの社会的背景を違う指標で表現することだ．序論の関連研究でも述べたように Twitter 上の影響力について分析する研究は数多く行われている．そうした研究の知見を生かせれば，ユーザの社会的背景を新しい方法で指標化し，そうしたものを特徴量として加えることができるかもしれない．このように社会的背景の指標か方法をより精査することで精度改善がもたらされる可能性があり今後取り組んでいきたい課題である．

参考文献

[1] M.Tanaka K.Ito, T.Ida. The persistence of moral suasion and economic incentives:field experimental evidence from energy demand, 2015.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013.

[3] Renaud Lambiotte Etienne Lefebvre Vincent D.Blondel, Jean-Loup Guillaume. Fast unfolding of communities in large networks, journal of statistical mechanics. *Theory and Experiment*, 2008.

[4] 家子直幸, 小林庸平, 松岡夏子, 西尾信治. エビデンスで変わる政策形成〜イギリスにおける「エビデンスに基づく政策」の動向、ランダム化比較試験による実証、及び日本への示唆〜, 2016. `http://www.murc.jp/thinktank/rc/politics/politics_detail/seiken_160212.pdf`.