ARG WI2 No.13, 2018
WI2-2018-26

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

久保田 修平[a]　大知 正直[b]　長濱 憲[†]　阪井 完二[†]
榊 剛史[c]　森 純一郎　坂田 一郎

東京大学工学系研究科　†株式会社　電通パブリックリレーションズ

a) *shu.kubota78@gmail.com*　b) *masanao.oochi@gmail.com*　c) *t.sakaki@hottolink.co.jp*

概要　近年，政策の現場では効率的に政策の企画立案や評価を行うため，客観的根拠に基づいて政策判断をしていこうとする機運が高まっている．本研究では，そうした政策分野への貢献を目指し，人々が関心を持つ社会的課題を定量的に可視化する試みとして，Twitter 上からの効率的な社会課題抽出とコミュニティ分割に基づくユーザコミュニティと社会課題の関係性に関する分析を行った．それにより，Twitter という単語や文法が崩れたノイジーなメディアからの社会課題抽出において，単語の係り受け情報が有効であることが示された．また，ユーザコミュニティと関心のある社会課題の関係性を分析することで，Twitter 上における社会課題を概観する新しいフレームワークを提案した．

キーワード　EBPM，Twitter，固有表現抽出，コミュニティ分割，社会課題

1　はじめに

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．不確実性が高まる現代において，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．

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．それにより，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等に関するデータが大量に蓄積されている．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 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．

自然言語処理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．前述の議論を踏まえれば，ソーシャルメディアに蓄積されている人々の社会的関心を固有表現抽出の技術を用いて抽出できれば，政策分野に対するエビデンスとして期待できる．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が短文で文法なども崩れたノイジーなものであることも多く，従来の手法を適用するだけでは十分とは言えないことが挙げられる．２つ目は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など社会的背景を考慮に入れられていないことだ．特に社会的な事柄では投稿を行なったユーザ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に関する知見はまだ少ない．

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し，前述した 2 つの課題を解決することを試みる．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で，各社会課題の社会的な重要度や関心の広がりなどを評価し社会課題の全体図を概観することができる．本研究ではここまでを最終的な目的としたい．

2　関連研究

本研究と同様に Twitter における固有表現抽出に関する分析を行った研究に Ritter らの研究 [1] がある．Ritter らは Twitter という短文で文法も崩れたノイジーな表現も多いメディアにおいて，表現の崩れを吸収するクラスタリングや人力でのタグ付けを行うことで品詞のタグ付けやチャンキング分析の精度を向上させ，それにより固有表現抽出の精度改善を実現した．

他にも Twitter から特定の情報を抽出する研究は数多く行われており，山本らは Twitter のデータから生活に関連する語の辞書を作成し，電車の遅延情報といった特定の地域に紐づく実生活情報を抽出する研究を行った [2]．また，Boettcher らの研究では，Twitter の投稿からある場所におけるイベントを抽出する研究を行っている [3]．ただし，これがソーシャルメディア上であるこ

とを加味した上で，ユーザの社会的背景等を考慮した抽出などを行なっているわけではない．

このように，社会課題に関して固有表現を抽出する研究についてはまだ知見が十分であるとは言えない．以上を踏まえ，本研究では，ソーシャルメディア上におけるノイジーな文章に対応し，ユーザの社会的背景などを考慮した固有表現抽出に関する実験を行い，考察を試みることにする．

3 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つはユーザのクラスタ分析及び，各クラスタと社会課題の関係性分析である．

3.1 社会課題抽出モデル

前節でも述べたように，本研究は社会課題に特化した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いくタスクとして考えられる．固有表現タグとして，本研究では IOB タグを使用する．IOB タグでは，固有表現の先頭の単語に B タグ，同一の固有表現で B タグに連続して繋がっている単語に I タグ，そのほかの単語に O タグをふっていく．このように各単語を IOB のタグに分類していくことで固有表現を文章中から抽出することが可能になる．

3.1.1 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．以下で本研究で利用する特徴量を具体的に述べていく．

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

3.1.2 品詞と文字種

本研究では，単語の特徴量を構成する際に，その単語の品詞と文字種を利用することにする．品詞に関しては，本研究で使用する MeCab という形態素解析エンジンに基づいて構成している．また，後者の文字種とは「ひらがな」や「カタカナ」といった文字の種類のことである．

3.1.3 単語の分散表現

本研究では Mikolov らの skip-gram Negative Sampling(SGNS) モデル [4] を用いて単語の分散表現を構成し，社会課題抽出モデルの特徴量として採用している．

skip-gram モデルは分布仮説に基づいて，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行うことでその単語の意味的な表現を獲得する．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である．

3.1.4 文脈表現

社会課題が現れる文脈をノイジーなデータ上で学習する際，社会課題に隣接する単語を利用するよりも，単語の係り受け情報を用いることでより直接的な形で社会課題の置かれている文脈を利用したほうが精度向上が期待できる．よって，本研究では前述の特徴量に加え，単語の係り受け情報を特徴量として用いる．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．文節はたいてい複数の単語で構成されるため，文節を構成する単語の分散表現の平均値を用いる．

3.1.5 ネットワーク特徴量

前述したように，発言ユーザの信頼性といった社会的背景に関する情報が，社会課題抽出に役立つことが期待される．本研究では，ユーザが引用しているメディアや会話しているユーザなどの情報がユーザの社会的背景を表すと考え，次のように特徴量を構成する．

Twitter 上においてユーザの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/される」「RT する/される」「メディアを引用する」の５つである．さらに，これらインタラクションに関して,「重複を許さないでどれだけ多くインタラクションがあったか」と「重複を許してどれだけ多くインタラクションがあったか」の２つの観点で分類し，特徴量を構成している．それぞれ，その人の考えの多様性や社会性，社会との関わりの量を表現していると考えられ，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表 1 に示す．

3.2 予測モデル

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「品詞・文字種・単語分散表現」のみを利用し，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．

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．

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)

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

3.3 ユーザクラスタリングと社会課題の分類

3.3.1 クラスターの抽出

ユーザをクラスタリングするためにメディアとユーザの会話ネットワークを利用する．本研究ではより興味関心の近い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，ユーザが引用する web 上の情報ソースも含めた，メディアとユーザの混合のネットワーク上でクラスタリングを行う．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はそのメディアとユーザ間に，またユーザ間にメンションがあった場合にその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．

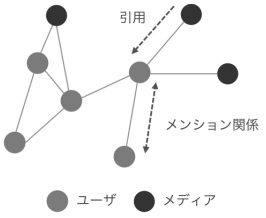

図 2 ネットワークの構成方法

このように構成したネットワークを Louvain 法 [5] 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である．

3.3.2 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，1 ユーザーのプロフィール文を 1 つの文書とみなして 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．

3.3.3 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．広く様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」，同様に「3. 一部の人にとって重要な社会課題」,「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ることにする．

	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．ツイート量に関しては単純なツイート数の log をとったものを用い，コミュニティの偏りの指標としては鳥海らの研究 [6] を参考にしエントロピーを採用する．さらに各軸を全体で標準化したスコアに変換することで 2 次元上に社会課題を整理することにする．ここで，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．

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

4 実験と結果

4.1 データセット

本研究では，2 種類のデータを利用する．社会課題の正解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．

4.2 社会課題抽出モデル

4.2.1 モデルの評価

モデルを評価する上で，特定の社会課題を含むデータをテストに利用し，含まない残りを学習データに利用する．学習データでは社会課題を含むツイートとなんの社会課題も含まないツイートを 1:1 で混ぜたツイート利用

し，テストデータに含まれる社会課題を識別できるかで精度の評価を行う．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 ツイート利用している．

4.2.2 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

4.3 ユーザクラスタリングと社会課題の分類

4.3.1 ユーザクラスタリングの結果

本研究では Louvain 法を用いて 21 のクラスターに分割を行なった．その結果の一部を表 4 に示す．

表 4 クラスター (一部抜粋)

cluster	ノード数	メディア	キーワード
0	2095578	soccer-king.jp, nikkans-ports.com	部, サッカー, 高校, 野球, バスケ, 選手
5	814859	www3.nhk.or.jp, sankei.com	原発, 政治, 日本, 反対, 憲法
8	590684	oricon.co.jp, ntv.co.jp	担,KinKi, 嵐, SMAP,NEWS

4.3.2 社会課題の分類

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のが図 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」，右下が「2. 社会課題の芽」，左上が「3. 一部の人にとって重要な社会課題」，左下が「4. 一部の人にとっての社会課題の芽」に対応している．「1. 大きな社会課題」として熱中症や過労死，また,「2. 社会課題の芽」として QOL(Quolity Of Life:生活の質) や既読無視や英語力といった言葉が得られた．

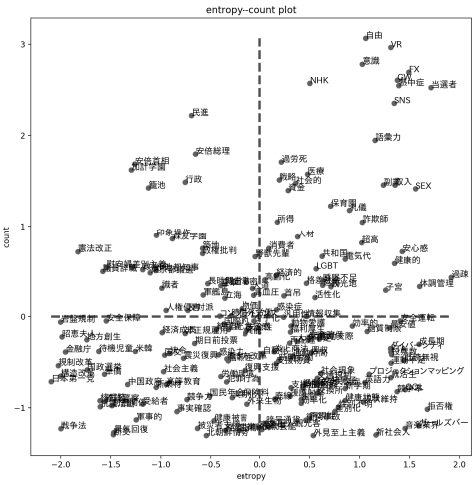

図 4 社会課題マッピング

5 まとめ

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い，Twitter というノイジーなメディア上での社会課題抽出において，係り受け情報が有用で，今回設定したネットワーク特徴量はあまり寄与しないことが示された．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．

結果から見られたようにユーザの信頼性に関する指標に改善の余地が存在している．社会的背景の指標化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] Ritter,A., Clark,S., Mausam and Etzioni,O.: Named Entity Recognition in Tweets:An Experimental Study, Conference on Empirical Methods in Natural Language Processing, 2011.

[2] 山本修平，佐藤哲司: Twitter からの実生活情報の抽出法の提案, DEIM Forum, 2012.

[3] Boettcher,A. and Lee,D.: EventRadar: A Real-Time Local Event Detection Scheme Using Twitter Stream, Green Computing and Communications, 2012.

[4] Mikolov,T., Chen,K., Corrado,G. and Dean,J.: Efficient Estimation of Word Representations in Vector Space, International Conference on Learning Representations, 2013.

[5] Blondel,V.D., Guillaume,J., Lambiotte,R. and Lefebvre,E.: Fast unfolding of communities in large networks, Journal of Statistical Mechanics, Journal of Statistical Mechanics: Theory and Experiment, 2008.

[6] 鳥海 不二夫, 榊 剛史: バースト現象におけるトピック分析, 情報処理学会, Vol. 58, No. 6, pp. 1287 - 1299, 2017.