# ソーシャルメディアの投稿は社会経済的地位のシグナルになるか？―サーベイ-ツイート統合データの機械学習モデルによる分析

水野誠[a] 瀧川裕貴[b]

**要約**

本研究では，ソーシャルメディアでの投稿が消費者の社会経済的地位（SES）に関するシグナルとして働く可能性を，同じ個人のサーベイデータと投稿データを統合して分析することで探求する．SES は経済学では所得・資産や学歴，社会学では職業威信，さらには文化資本との関係で捉えられることが多い．本研究では別の視点から，産業社会の変化を反映した職業のクリエイティブネスや社会的必要性といった項目も考慮する．分析においては第１に，ソーシャルメディアでの投稿内容（潜在トピックや特定ブランドへの言及）に，ユーザの SES に基づく差異が見られるかどうかを分析する．これが認められれば，投稿が SES のシグナルとなる必要条件が満たされる．第２に，ソーシャルメディア上の投稿内容（使用された語）から SES を予測できるかを分析する．一般の人々にもそうした予測が可能だとしたら，投稿が SES のシグナルとなる十分条件が満たされるだろう．これらの分析により，現代の消費行動において SES の果たす役割について探求する．

JEL 分類番号： Z13, M31, D12

キーワード：ソーシャルメディア，社会経済的地位，消費，機械学習

[a] 明治大学商学部 makoto@meiji.ac.jp

[b] 東京大学文学部・大学院人文社会系研究科 takikawa@l.u-tokyo.ac.jp

## 1. イントロダクション

消費者の社会経済的地位(SES)が彼らの消費水準に影響することは，SES を所得や資産で測る限り，経済学的には自明といってよい．個別の財やブランドの消費に対する SES の影響には，経済学には Giffen 財や Veblen 財（Veblen 1899）といった古典的な議論があり，社会学における文化資本に関する議論も重要である（Bourdieu 1979）．

本研究は，これらの研究を継承しつつ，分析対象をソーシャルメディア上のコミュニケーションに設定する．そこでユーザはさまざまな投稿を行っており，消費や余暇活動に言及することも多い．そのとき SES のシグナリングが行われている可能性を，自然に大規模に集積されるデータ（デジタルトレース）を用いて経験的に検証することを目指す．そのための準備作業として，500 近い職業について独自の SES スコアを算出し，Twitter ユーザに対するサーベイで職業を把握し，彼らの SES スコアを推計する．

ソーシャルメディア上の投稿がその発信者の SES のシグナルとなるための必要条件は，投稿の内容（特定の単語の使用や特定のトピックへの言及）に SES に基づく差異があることであろう．そこで本研究ではまず，他の要因をコントロールした上で，SES の違いが投稿内容に有意な差を生み出しているどうかを，投稿の背後にある潜在的なトピック，また特定のブランドに関する言及を通して分析する．

他方，投稿が SES のシグナルとなる十分条件は，投稿内容から SES を予測できる可能性があるかどうかで検証されると考える．そのためさらに彼らの投稿（ツイート）を大量に収集し，両者を統合する（Stier, et al. 2020）．そこに機械学習の手法を適用して，ツイートのテキストデータから SES スコア等の個人属性を予測する．そうした予測を行った研究はすでに多く存在するが（He and Tsvetkova 2022; Sloan et al. 2015; Yo and Sasahara 2017 など），本研究は，既知であるユーザの SES を正しく識別できるかどうかに関心を絞っている点で，それらの研究とは違う[1]．すなわち，人間にとっての予測可能性の検証のため，機械に予測タスクをさせている．

## 2. データ収集

### 2.1. 社会経済的地位のスコア化

本研究では SES を単に所得や資産で測定される以上のものと考える． SES スコアの測定項目として，社会学における階層研究で頻繁に用いられてきた職業威信（プレステージ）以外に，名声，権力，収入，学歴，文化，クリエイティブネス，社会的必要性などに関する社会的に共有される知覚を考慮する．このうち，職業のクリエイティブネスはクリエイティ

[1] Grimmer, Robert and Steward 2021 はそうした分析を疑似予測法と呼んでいる．

ブ・クラス論（Florida 2002, 2012），社会的必要性はブルシットジョブ論（Graeber 2018）やコロナ禍でのエッセンシャルワーカーへの関心を反映させたものである．

職業としては日本版 O-NET のサイトに掲載された 441 の職業に，現代的な職業を追加し，各中分類の「その他」の選択肢を補足することで，約 500 の選択肢を用意した．スコアリングはウェブ調査とクラウドソーシングを利用して複数回実施された．これらの調査は１つの職業に約 100 名の評価が得られるよう設計されている（20 歳以上の日本在住の有職者から性別・年齢に偏りがないよう抽出される）．

各項目は５点尺度で評価されている．いくつかの項目間には強い相関関係があるため，職業威信，クリエイティブ，社会的必要性の３項目を今後の分析で用いることにする．職業威信の高い職業は法曹・医療関係の専門家が多く，クリエイティブな職業としてはデザインに関わる職業が多い．社会的必要性が高いのは看護師，救急救命士などのいわゆるエッセンシャルワーカーが多い．なお，職業威信とクリエイティブネスには正の相関がある．

### 2.2. Twitter ユーザのウェブ調査とツイート収集

Twitter ユーザーを対象にウェブ調査を実施し 5,196 人の回答を得た．対象者の選定条件は 20 歳以上の日本在住者で，月に 2 回以上 Twitter に投稿していることで，性・年齢・特定の職業に関する割付も行われた．回答者については，職業以外に，最終学歴，個人・世帯収入，デモグラフィクス（年齢，性別，家族構成，居住地域）に関する情報を入手している（多くがパネル属性として調査会社から提供された）．

ウェブ調査で回答された職業は，2.1 で得られたスコア化の手続きに基づき，３項目の SES スコアに変換された．次に各回答者から許可を得て取得した Twitter アカウントに基づき，Twitter API を用いて過去の投稿（ツイート）を収集した.分析可能なデータは 3,930 人の回答者について得られている．

## 3. データ分析

### 3.1. ツイートのトピック・モデル分析

収集されたツイートに対してトピック・モデルによる分析を行う．１人の Twitter ユーザの全投稿を１つの文書として LDA (Latent Dirichlet Allocation)を適用し，15 の潜在トピックからなるモデルを推定する（トピック数は試行錯誤により選択された）．次に各トピックと強く関連する単語を各トピックの解釈を行う（紙幅の都合で詳細は省略する）．

各トピックに対する各回答者の確率を SES スコアやデモグラフィクスに回帰させた結果が，表 1 に示されている．ほぼすべてのトピックについて，性別・年齢による差が見られる．また約半数のトピックにおいて，職業威信あるいはクリエイティブネスによる差が見ら

れる．ただし，いずれの場合も決定係数はかなり低い．要約すると，性別・年齢によるトピックの違いは全般的であり，SES スコアによる違いも一定程度見られる．

表 1　各トピックと個人属性の関係に関する線形回帰

| | 1<br>懸賞 | 2<br>英語など | 3<br>コロナ | 4<br>懸賞 | 5<br>食事 | 6<br>交流 | 7<br>ペット | 8<br>配信 | 9<br>絵文字 | 10<br>懸賞 | 11<br>スポーツ観戦 | 12<br>メール連絡 | 13<br>懸賞 | 14<br>日常 | 15<br>推し |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 職業威信 | -.003 | .027 | .058 | .025 | .023 | -.062 | -.010 | -.067 | -.059 | -.072 | .040 | -.006 | -.043 | .051 | .012 |
| クリエイティブ | .054 | -.021 | -.001 | -.067 | .029 | .071 | .000 | .054 | .012 | -.008 | -.058 | -.009 | -.044 | .019 | .031 |
| 社会的必要 | .002 | -.026 | -.028 | .043 | -.036 | -.019 | -.016 | -.018 | .040 | .009 | -.030 | .019 | .047 | -.032 | .004 |
| 性別(女) | -.127 | -.084 | -.135 | .036 | -.053 | -.011 | -.002 | .092 | .188 | -.112 | -.263 | -.040 | -.008 | .072 | .176 |
| 年齢 | .111 | .045 | .294 | .121 | .250 | -.115 | .065 | -.014 | -.068 | -.129 | -.114 | .043 | .019 | -.090 | -.350 |
| 既婚 | .027 | .006 | -.053 | .069 | .033 | -.031 | -.003 | -.091 | -.007 | .024 | -.055 | .003 | -.014 | .039 | .014 |
| 子あり | .021 | .026 | -.032 | .025 | -.045 | .026 | -.040 | -.050 | -.022 | .064 | .027 | .031 | .086 | -.053 | -.032 |
| 北海道・東北 | .022 | .004 | -.023 | .008 | .003 | .026 | .024 | -.004 | .015 | .041 | .008 | -.003 | .006 | -.021 | -.035 |
| 中部 | .021 | .010 | -.019 | .040 | .000 | .007 | -.013 | -.064 | .025 | .042 | -.011 | -.017 | .023 | -.053 | .015 |
| 関西 | .012 | -.009 | -.017 | .047 | -.025 | .064 | -.005 | -.042 | .002 | .005 | -.012 | .022 | .037 | -.062 | .007 |
| 中部・四国 | .010 | .039 | -.027 | .024 | -.017 | -.007 | .028 | -.065 | .012 | .017 | -.015 | -.007 | .022 | -.004 | -.002 |
| 九州・沖縄 | .010 | -.004 | .018 | .034 | -.009 | .021 | .002 | -.035 | .024 | .022 | -.012 | -.011 | -.004 | -.036 | -.005 |
| $R^2$ | .096 | .024 | .073 | .032 | .061 | .047 | .013 | .041 | .110 | .058 | .140 | .020 | .025 | .024 | .150 |

＊数値は最終行を除き標準化回帰係数．$p < .01$ のケースを矩形で囲った．

### 3.2 ブランドへの言及の分析

各回答者のツイートにおけるブランドへの言及の有無を，SES，デモグラフィクス，そして投稿文字数で予測するロジスティック回帰分析を行う．対象ブランドは，丸の内ブランドフォーラムが実施したブランド想起に関する調査に基づき選択された．ブランドへの言及の有無の定義は，その個人の全投稿に 10 回以上表記ゆれを含むブランド名が出現することである（投稿数の違いは全文字数の変数によってコントロールされている）．分析結果から，トピックの場合ほどではないが，性別・年齢は多くのブランドへの言及と関係することが示された（詳細は省略）．他方，SES スコアが有意な関係にあるブランドは限られる．職業威信と正の関係があるのは「Yahoo!」「日本経済新聞」「東京大学」の 3 ブランドである．

### 3.3 ツイートから SES の予測

最後に，ツイート内容（単語）から職業威信や所得など SES に関わる変数を予測する．機械学習を用いた予測が正確なら，消費者もまたはそのツイートだけから投稿者の SES を推測できる可能性が高いと考える（一般に人間の予測能力のほうが高いと想定）．サポートベクターマシンやランダムフォレストなどの教師付き機械学習の手法をいくつか試したのち，この問題について最も予測性能が高かった L2 正則化ロジスティック回帰を採用した．

なお，比較のため SES スコアだけでなく，性別・年齢のようなデモグラフィクスも予測対象に加えた（ただし連続変数は二値に離散化された）．予測の成否をデータ内のその属性の比率に基づくナイーブな予測以上に正確かどうかで評価すると，成功したといえるのは，

年齢（40 歳以上か未満か）と性別だけであり，SES の各種変数についてはそうではなかった．表 2 に，性別と年齢に関して，予測にとって重要な特徴量となる語を挙げておく．

表 2　性別・年齢を予測する重要な特徴量

| | 男性 | 女性 | | 男性 | 女性 | | 40歳未満 | 40歳以上 | | 40歳未満 | 40歳以上 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 俺 | くん | 11 | pepsi | 肌 | 1 | 僕 | 応援 | 11 | きた | 新潟 |
| 2 | 僕 | ハート | 12 | チャンス | セット | 2 | めちゃくちゃ | 不機嫌 | 12 | 記録 | 投稿 |
| 3 | ラーメン | 嬉しい | 13 | 勝つ | 大好き | 3 | まじ | 娘 | 13 | pokemon | 艸 |
| 4 | 車 | わたし | 14 | 食う | gt | 4 | やばい | ちょっと | 14 | やつ | 怒り |
| 5 | 投資 | 素敵 | 15 | ヨドバシ | 商品 | 5 | 俺 | 暑い | 15 | くらい | research |
| 6 | いく | 先生 | 16 | っす | 病院 | 6 | 可愛い | 美味しい | 16 | 保育園 | 今朝 |
| 7 | ビール | 旦那 | 17 | チェキ | ひよこ | 7 | 分かる | 下さる | 17 | チーズ | 日 |
| 8 | jal | あたし | 18 | 期待 | 札幌 | 8 | fgo | 見える | 18 | 終わる | 楽天スーパーポイント |
| 9 | 無い | 合掌 | 19 | やはり | 大泣き | 9 | キラキラ | 息子 | 19 | 行く | ヽ |
| 10 | 乃木坂46 | 星野源 | 20 | アニメ | 涙 | 10 | いく | 汗 | 20 | ほしい | 子 |

4. 議論

以上の分析結果を要約すると，Twitter ユーザは， SES の差異に応じて投稿におけるトピックが変わり，特定のブランドへの言及度合いが変わる．したがって，SES がこれらの情報発信に影響している可能性は十分ある．しかし，投稿内容だけから SES を予測することに機械学習が失敗したため，人間にとってもそれは可能だという証拠は得られなかった．したがって，現段階ではツイートが SES のシグナルとして働くとはいい難い．

このような結果になった背景には，以下のような可能性があると考えられる：

(1) 性別や年齢に比べ，SES の差を示す語彙が（特に日本語に）乏しい

(2) 最近，SES をあからさまにシグナルすることは好まれない（Berger and Ward 2010）

(3) 日本では文化的な社交の場で SES の差を誇示することは伝統的に避けられがちで（Ikegami 2005），それがソーシャルメディアでも起きた

(4) SES のシグナリングは Twitter 以外のメディア（Instagram 等）で起きやすい

(5) Twitter ユーザ内でも，特定の個人が特殊な状況で SES のシグナリングを行っている

(1)は少なくとも言語によるシグナリングの可能性を否定するが，(4)で述べた画像系のソーシャルメディアでのシグナリングの可能性を否定しない．(2)は Veblen の古典的な誇示的消費（conspicuous consumption）が現在も存在していることには懐疑的だが，微妙で複雑なシグナリングについて問題提起している．そこでは，そのシグナルを読み取る能力が特定の文化資本に依存することを示唆している．そこから，(5)でいう異質性の扱いについて，１つの方向性が提案されているといえよう．(3)は日本社会の歴史的背景に関するより深い議論が要求される．いずれにしろ今後いっそう研究を進めることが望まれる．

引用文献

Berger, J., and M. Ward, 2010. Subtle signals of inconspicuous consumption. Journal of Cconsumer Research, 37(4), 555-569.

Bourdieu, P., 1979. La distinction, critique sociale du jugement. Les Editions de Minuit, Paris, France. （石井洋二郎訳, 2020, ディスタンクシオン：社会的判断力批判. 藤原書店, 東京）

Florida, R., 2002, The Rise of the Creative Class. Basic books.（井口典夫訳, 2008, クリエイティブ資本論. ダイヤモンド社, 東京）

Florida, R. 2012, The Rise of the Creative Class, Revisited. Basic books.（井口典夫訳, 2014, 新 クリエイティブ資本論, ダイヤモンド社, 東京）

Graeber, D. 2018, Bullshit jobs: A theory. Simon & Schuster. (酒井隆史・芳賀達彦・森田和樹訳, 2020, ブルシット・ジョブ クソどうでもいい仕事の理論. 岩波書店, 東京)

Grimmer, J., M.E. Roberts, and B.M. Stewart, 2022. Machine learning for social science: an agnostic approach. Annual Review of Political Science, 24, 395–419.

He, Y. and M. Tsvetkova, 2022. A Method for estimating individual socioeconomic status of twitter users. arXiv:2203.11636.

Ikegami, E., 2005. Bonds of civility: Aesthetic networks and the political origins of Japanese culture, Cambridge University Press, UK. (池上英子, 2005. 美と礼節の絆：日本における交際文化の政治的起源, NTT 出版, 東京)

Yo, T. and K. Sasahara, 2017. Inference of personal attributes from tweets using machine learning. IEEE International Conference on Big Data, 3168–3174.

Sloan, L., J. Morgan, P. Burnap, and M. Williams, 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PLOS ONE, 10(3), e0115545.

Stier, S., J. Breuer, P. Siegers, and K. Thorson, 2020. Integrating survey data and digital trace data: Key issues in developing an emerging field. Social Science Computer Review, 38(5), 503–516.

Veblen, T., 1899. The Theory of the Leisure Class. MacMillan Company, New York, US. (高哲男訳, 2015. 有閑階級の理論. 講談社, 東京)