

ツイートが運ぶ多様な個人属性の定量化

Quantification of Diverse Personal Attributes in Tweets

余 岳 *1

Take Yo

佐治 礼仁 *2

Ayahito Saji

笹原 和俊 *1*3

Kazutoshi Sasahara

*1名古屋大学大学院情報学研究科
Graduate School of Informatics, Nagoya University

*2名古屋大学情報学部
School of Informatics, Nagoya University

*2名古屋大学大学院情報学研究科, JST さきがけ
Graduate School of Informatics, Nagoya University, JST PRESTO

We studied personal attributes represented in tweets, such as gender, occupation, and age groups. First, we examined how much these basic attributes can be predicted from the texts of tweets, each of which was vectorized by a word2vec-based method for machine learning. The results showed that machine learning algorithms can predict all three attributes with 60-70% accuracy. We also confirmed that differences in word usage between males and females (related to semantic differences) affect the predictive accuracy of gender. Furthermore, we quantified other personal attributes, such as Big 5 and values, using IBM Personality Insights.

1. はじめに

近年、情報のハブかつ意見交換のプラットフォームであるソーシャルメディアから、人間行動に関するビッグデータを収集できるようになった。このような時代背景から、計算社会科学(Computational Social Science)と呼ばれる新しい学際科学が誕生し、現在盛んに研究が行われている[Lazer 09]。例えば、Twitterでは「ゆるいつながり」による多様なコミュニケーションを観測することができ、ソーシャルデータの入手が容易なため、人間行動や社会現象の理解に利用されている[Sasahara 13, Takeichi 15]。

ソーシャルデータを用いた研究の中でも、個人属性の推定に関する研究は活発に行われている。ソーシャルデータに含まれる大量の言語表現には、パーソナリティーを反映した情報が潜在していると考えられる[Pennebaker 15]。Schwartzらは、Facebookの投稿に含まれる単語とトピックの使用頻度に基づいて辞書を作成し、性格、性別、年齢を判別できることを示した[Schwartz 13]。Kosinskiらは、Facebookの「いいね！」の頻度に基づいて、線形回帰などの比較的簡単な方法でも個人属性を推定できることを示した[Kosinski 13]。Liuたちは、中国のソーシャルメディアWeiboの投稿に基づいて、autoencoderを学習させ、投稿者のビッグ・ファイブと呼ばれる人間の性格因子(神経症傾向、外向性、経験への開放性、協調性、誠実性)を推定できることを示した[Liu 16]。これらのように、ソーシャルデータから個人情報を数値化・推定する手法は、多くの研究者が取り組んでいるものの、まだ確立した一般性の高い方法は存在しない。

本研究の目的は、ソーシャルデータのテキストのみから個人属性をどの程度推定できるかに関するベースラインの知見を得し、その理由を調査することである。そこで、Twitterのデータを用いて、機械学習のアルゴリズムによって性別、職業、年齢層の3つの個人情報を予測する実験を行った。さらに、性別を例とし、男女の共通単語の予測精度への影響を調べた。

連絡先: 余 岳, 464-8601 名古屋市千種区不老町

名古屋大学大学院情報学研究科, yotake1987@nagoya-u.jp

2. 方法

2.1 データ収集

本研究では予測モデルを構築するために、Web検索で性別と職業が確認でき、投稿されたツイート数が3000以上のアクティビティアカウントを120人特定した。また、これらのアカウントの年齢をWeb検索で可能な限り特定した(年齢は非公開な場合が少なくない)。これらのアカウントのうち、100人をトレーニング用、残り20人をテスト用にした。

120人のアカウントの特徴は、次のようにまとめられる。男性と女性のアカウント数は等しく、偏りがない。職業は10種類を選択し、それぞれの職業に属するアカウント数は等しく、偏りがない。年齢層は、1980年以後の生まれ(デジタルネイティブ)か、それ以前の生まれ(デジタル移民)の2種類にした。

次に、これらのアカウントからTwitterの公式APIを用いて、リツイートや返信を含むユーザータイムラインを可能な限り収集した。その結果、トレーニング用のアカウントからは314382個、テスト用のアカウントからは64027個のツイートが収集された。

2.2 データ処理

次の手順でデータ処理を行った。収集したツイートを、日本語形態素解析ツールMeCabと日本語辞書NEologdを用いて分かち書きの処理をした。分かち書きしたツイートから、長さが4未満の情報量が少ないツイートを削除した。クリーニング後に残ったツイートは、トレーニング用が312169個、テスト用が63454個である。トレーニング用の全ツイートをコーパスとして(全単語数は11308535個、異なり語は395491個)、word2vec[Mikolov 13]を用いて(window sizeは5、イテレーション回数は20)、全ての異なり語を単語ベクトルの辞書に変換した。

各ツイートのベクトルは、そこに含まれる単語のベクトルの平均値を使用した。単体のツイートに含まれる情報量は少ないため、複数のツイートを束にして学習させる方が精度が向上する可能性があると考えられる。そこで、複数のツイートを束にしたものツイートブロックと呼び、そこに含まれる各ツイ

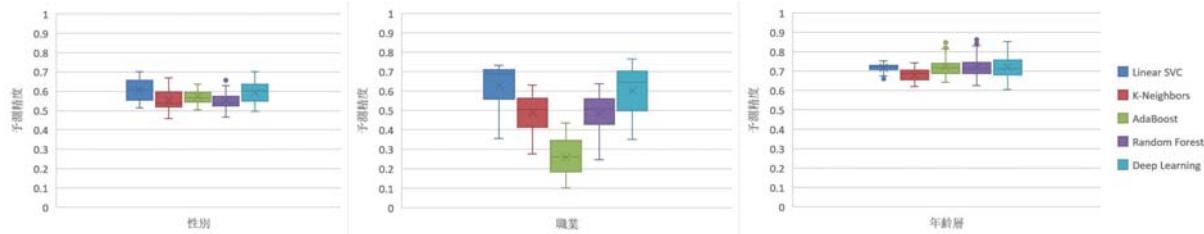


図 1: 各課題における予測精度の分布

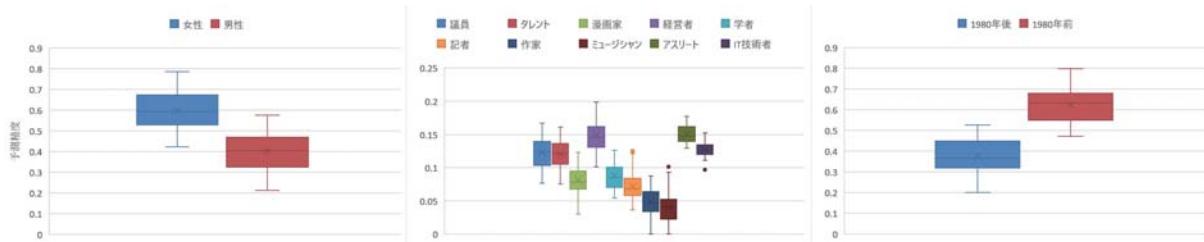


図 2: 個人属性における各項目の予測のしやすさ

トをベクトル的に平均化したものツイートブロックのベクトルとして扱った。

2.3 基本的な個人属性の予測

以上の処理から得られたツイートブロックのベクトル値を入力として、機械学習を用いてモデルを構築し、個人属性の予測精度を調べる。使用した学習アルゴリズムは、Linear Support Vector Classification (Linear SVC), K-Neighbors, AdaBoost, Random Forest の 4 つと深層学習である。

2.4 その他の個人属性の定量化

本研究の 120 人の Twitter のアカウントに対し、IBM Watson の Personality Insights を使って得られたデータを定量化した。Personality Insights の公式 API を用い、各アカウントにおけるビッグ・ファイブ（5 項目）、ニーズ（12 項目）、バリュー（5 項目）の合計 22 項目の個人属性のデータを取得した。本研究では、特に職業に注目して、22 項目の個人属性を 22 次元のベクトルとして各アカウントを特徴付ける。各職業ごとに 12 名のアカウントの平均値をとり、その職業の特徴ベクトルにする。そして、主成分分析の手法で各職業の 2 次元の空間配置を観察する。

2.5 個人属性の予測に影響する要因の分析

個人属性の予測が成功する要因を調べるために、次のようなデータ分析を行う。例えば、ツイート・コーパスに存在している男女が共通して使用する単語は、性別の予測するときに重要な役割を果たすと推測される。ここでは、性別を例としてデータを分析し、この考えを検証する。トレーニング用のツイート・コーパスを性別ごとに分け、前節と同様の処理で単語の分散表現を作り、男女ごとにそれぞれ単語ベクトルの辞書を作成した。男性の単語ベクトル辞書（単語数 237141）と女性の単語ベクトル辞書（単語数 241645）に基づき、共通単語の集合（単語数 83295）を抽出した。そして、男女の共通単語において、コサイン距離を測った。

予測精度への影響を調べるために、トレーニング用のコーパスから（単語数が 376577）から、重み W （コサイン距離と出現頻度の積）の大きい順に従って単語を取り除き、新しい単語

ベクトル辞書を作成した。この辞書を用い、トレーニング用の 100 人とテスト用の 20 人のすべてのツイートをベクトル化した。また、非共通単語をランダムに取り除き、同様の処理をして、比較対象とした。

3. 結果

3.1 基本的な個人属性の予測

図 1 は、word2vec の埋め込み次元 N とツイートブロックのサイズ L の全ての組み合わせにおいて、3 つの個人属性の推定における精度の分布を示したものである。年齢層は、どの学習アルゴリズムにおいても平均して 70% 程度の予測精度を示していることから、ツイートのテキストのみから年齢層を推定することは、他の 2 つと比べて容易であることがわかる。また今回の実験では、Linear SVC と深層学習は他のアルゴリズムと比べて、安定して高い精度が得られることがわかった。

3.2 個人属性の項目ごとの予測

性別、職業、年齢層の 3 つの個人属性には、予測しやすい（しにくく）項目があると考えられる。それを調べたのが図 2 で、深層学習（全ての N と L の組み合わせ）によって予測が成功した項目の内訳を示している。性別の場合には、男性より女性の方が予測しやすい。職業の場合には、経営者とアスリートは予測しやすく、作家とミュージシャンは予測しにくい。年齢層の場合には、デジタルネイティブよりデジタル移民の方が予測しやすいことがわかった。

3.3 その他の個人属性の定量化

Personality Insights の結果に基づき、22 次元で表現した各職業の個人属性（平均値）を、主成分分析で二次元にマップしたのが図 3 である。これを見ると、学者と記者、タレントとミュージシャンは個人属性が似ていることがわかる。一方、IT 技術者とアスリート、漫画家と経営者は、他の職業とは個人属性が似ていないことが読み取れる。

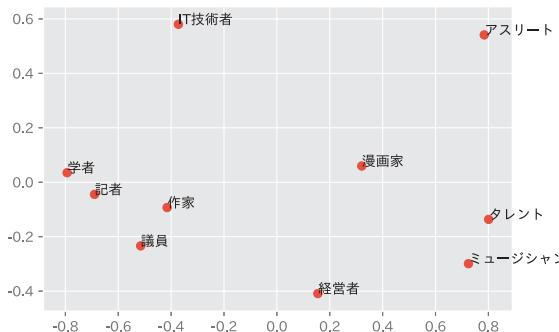


図 3: Personality Insights の個人属性のデータに基づく職業の類似度

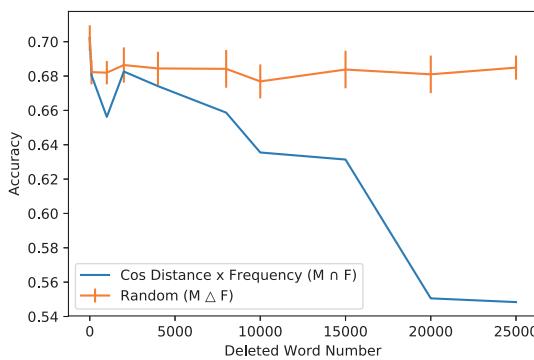


図 4: 男女の共通単語と非共通単語を取り除いた際の予測精度への影響

3.4 性別予測に影響する要因の分析

トレーニング用のツイート・コーパスから、重み W の大きい順に共通単語を 1 つずつ取り除いて、性別の予測精度を調べた。同様に、非共通単語をランダムに 1 つずつ取り除いて予測精度（10 回の実験結果による平均値）を比較した。ここでは、深層学習のアルゴリズムを用い、前の実験結果による $N=500$, $L=50$ という性別予測に最適なパラメーターで行った。その結果が図 4 である。取り除かれた共通単語の個数を増やすにつれて、予測精度は最初の 70% から 55% に大幅に下がった。一方、比較対象の非共通単語の場合は、取り除かれた単語数を増やしても予測精度がずっと 68% 程度に安定している。

4. まとめ

本研究では、まず、ツイートのテキストのみから性別、職業、年齢層という 3 つの個人属性を予測する実験を機械学習を用いて行い、ベースラインの結果を得た。

予測精度を比べると、年齢層（デジタルネイティブか否か）は、性別や職業よりもアルゴリズム的に推定しやすいことが示された。これは、言葉の選び方や使い方が年齢によって異なるからではないかと考えられる。テキストデータのみから、60 ~ 70% の精度で個人属性を推定できるということは注目に値する。ツイートから個人属性を予測する場合、単体のツイートよりも複数のツイートをまとめてブロックとして使用した方が精度が顕著に上がった。この結果は、ツイートの束をまとめてベ

クトル化したものが、個人属性の推定に必要な情報を比較的低次元で表現していることを示唆している。換言すると、テキストのみから個人属性を推定するためには、ある程度の量のデータが必要である。Twitter の場合は、60% 程度の予測精度を得るために、50 個程度がその下限だということになる。その他の個人属性の定量化から、個人属性が類似する職業や、逆に類似しない職業があることがわかった。

個人属性を予測できる理由を調べたところ、男女が共通して頻繁に使用する単語の中で意味が異なる単語は、性別の予測精度に大きい影響を与えることが示された。共通単語の場合は、取り除かれた単語数が増えていくにつれ、予測精度が著しく下がった。さらに、取り除かれた単語数は 15000 から 20000 の間にある時、最も予測精度が下がった。これは、一部の共通単語が他より予測精度に強い影響を与えていることを示唆している。非共通単語の場合は、取り除かれた単語数が増えても予測精度が大きく変動することがなかった。これらの結果によると、性別を予測する場合は、非共通単語と比べて共通単語が予測精度に大きく影響することが分かった。今後、性別以外の個人属性でも同様の調査をする予定である。

本研究の技術が確立すれば、ソーシャルデータから個人属性を精度良く推定することが可能になり、計算社会科学の分析やマーケティングへの応用など幅広い利用が期待できる。

謝辞

本研究は JSPS 科研費 (JP16K16112, JP15H03446, JP17H06383JST), JST さきがけ (JPMJPR16D6), JST CREST(JPMJCR17A4) の助成を受けたものです。

参考文献

- [Kosinski 13] Kosinski, M., Stillwell, D., and Graepel, T.: Private traits and attributes are predictable from digital records of human behavior, *Proceedings of the National Academy of Sciences*, Vol. 110, No. 15, pp. 5802–5805 (2013)
- [Lazer 09] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Alstyne, M.: Computational Social Science, *Science*, Vol. 323, No. 5915, pp. 721–723 (2009)
- [Liu 16] Liu, X. and Zhu, T.: Deep learning for constructing microblog behavior representation to identify social media user's personality, *PeerJ Computer Science*, Vol. 2, p. e81 (2016)
- [Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013)
- [Pennebaker 15] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J.: The Development and Psychometric Properties of LIWC2007 (2015)
- [Sasahara 13] Sasahara, K., Hirata, Y., Toyoda, M., Kitsuregawa, M., and Aihara, K.: Quantifying collective attention from tweet stream, *PLoS ONE*, Vol. 8, No. 4, p. e61823 (2013)
- [Schwartz 13] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al.: Personality, gender,

and age in the language of social media: The open-vocabulary approach, *PLoS ONE*, Vol. 8, No. 9, p. e73791 (2013)

[Takeichi 15] Takeichi, Y., Sasahara, K., Suzuki, R., and Arita, T.: Concurrent Bursty Behavior of Social Sensors in Sporting Events, *PLoS ONE*, Vol. 10, No. 12, pp. e0144646–13 (2015)