E-005

# 主観的幸福度のリアルタイム推定のための Google Trends および Twitter データの併用アプローチ
# A Combined Approach of Google Trends and Twitter Data for Real-time Estimation of Subjective Well-being

佐伯 凌 [1]　　湯浅 晃 [2]
Ryo Saeki　　Akira Yuasa

## 1. はじめに

主観的幸福度(SWB)の調査は政府が地域住民の状態の評価や行動予測を行うのに役立ち，また政策立案において政策や危機に対する反応を理解するためにも有用である．しかし現状の SWB 調査においては統計データや一部アンケート調査が用いられており実施頻度やリアルタイム性に課題がある．

この課題に対して SNS データを用いた検証が進んでおり，World Happiness Report2023 にて Kjell[1]らは，SNS から幸福度を評価するアプローチは成熟しつつありデータの収集方法の改善と，辞書ベースから大規模言語モデルへの洗練による精度の向上により従来よりも費用対効果が高く正確で堅牢なものとして期待できると述べている．

本稿におけるタスクを説明する．目的変数は SWB で 0 から 1 の値をとり，各種データから抽出される特徴量をもとにした回帰タスクとして定義される．

既存研究において Twitter または Google Trends データを用いた報告は存在するが，本稿の新規性はそれらのデータを併用し，かつ 2 段階のモデリングを行っている点にある．Google Trends データの主な内容は検索クエリであるため SWB に関わる人々の生活における生の声を捉えておらず，また検索量のボリュームを絶対値で取ることができず，質と量の点で問題がある．この問題に対しては Twitter データを用いることで解決できると考えられる．しかし，複数のツイートをもとに特徴量を得る場合，例えば各ツイートの単語分散表現の平均ベクトルを取る方法があるが，これを回帰モデルの入力とする場合，特徴量の寄与度に関する解釈性に問題が生じる．SWB に影響する要因として一般に家庭，教育，生活など様々なものがあり，各要因の SWB への寄与度を得ることは有用である．しかし，例えば BERT を用いる場合，説明変数は 768 次元のベクトルとなるが，各次元は解釈可能な意味を持たないため，要因の寄与に関する解釈ができない．

そこで提案手法では，まず 1 段目のモデリングとして SWB の要因ごとに状態を推定する．例えば，家庭，教育，生活などに関してネガティブまたはポジティブかを-1 から 1 のスコアで推定し，この出力結果を特徴量として，2 段目でその他のデータから得られる特徴量と併せて ElasticNet による回帰モデルの入力とする．

2019-2021 年の期間，6 カ国における主観的幸福度の推定において，ベースラインとした経済指標のみを使用した手法を上回る推定精度が得られた．考察の結果，経済指標には表現されない家庭，教育，生活等の状態をあらわす特徴量が得られたためと考えられ，主観的幸福度の推定におけるこれらのデータの有効性を確認した．

## 2. 関連研究

Carpi[2]らは，Twitter データに基づいた日本の主観的幸福度指標である SWB-J 指数を初めて提示した．この指数は主観的幸福度の 8 つの側面で構成されており，人間による教師付き感情分析を使用して Twitter データに基づいて推定される．さらに構造方程式モデルを用いて経済および健康状態と幸福度の潜在変数との関係を調査した．

Algan[3] らは Google Trends から取得した単語の検索量に基づいて米国の幸福度を推定するモデルを構築した．

テキストからの筆者の心理推定に関して，Mairesse[4]らは文章の特徴に筆者の性格との関連性が見出せると報告し，那須川[5]らはカテゴリ辞書と分散表現を用いた日本語筆者の BigFive に基づく性格推定手法を提案している．Mori[6]らは SNS から投稿者の社会性や健康状態などのパーソナリティを Twitter の利用状況や投稿内容等の情報から推定し，その有効性を示している．

性格推定タスクのための大規模言語モデルのチューニング手法として，著者らの研究グループにおける Paolo[7]は，データがクラスと極性のマルチレベル構造を持つ場合に有効な学習方法を提案している．そこでは学習において SentenceBERT を用いた 2 段階の学習アプローチの有効性が報告されている．

## 3. データセット

本研究では，目的変数に OECD が公開する Better Life Index を使用し，説明変数としては，マクロ経済指標をベースラインとして用い，Google Trends データおよび Twitter データとの様々な組み合わせ条件における比較実験を行った．それぞれのデータセットを以下で説明する．

**Better Life Index**

OECD が公開している Better Life Index の Subjective Well-being-Negative affect balance 指標を使用した．本指標は Gallup 世界調査をもとに各国の主観的ウェルビーイングを算出したものである．

**マクロ経済指標(MFE)**

経済指標として一人当たりの GDP，労働力率，インフレ率を使用した．

**GoogleTrends データ**

Google Trends は Google での各キーワードの検索量の遷移を確認できるサービスである．クエリとして検索キーワードと検索トピックを用いる方法があるが，検索キーワードは言語に依存するため検索トピックで収集する方が適切である．そのため本研究では検索トピックでデータを収集した．ただ各クエリで 5 つまでしか検索トピックを指定できず，また，クエリ間で検索量の絶対評価ができない点が課題であるが，クエリ毎に共通のダミーの検索トピックを指定し，その検索量で対象トピックの検索量を除算することで正規化を行った．

**Twitter データ**

学習データはハッシュタグを使用してあらかじめ生活満足度の各要因に関して，ポジティブまたはネガティブ

1 株式会社アイヴィス　IVIS, Inc.
2 株式会社 NTT データ　NTT DATA Corporation

な内容を表すキーワードを人手によりいくつか選定し，当該ハッシュタグを含むツイートを収集した．このようにしてツイートに対して各要因の正負のラベルが付与された弱教師ありデータセットを構築した．

推論データは時刻，場所による絞り込みを行いツイートデータを収集した．時刻指定は 2010-2021 年までの間で年毎に指定した．時間帯，季節の影響をなくすために，時間帯，月を満遍なく収集．場所指定は各国の首都から半径 10km 以内の位置情報を持つツイートを収集した．

## 4. 手法

提案手法では 2 段階のモデリングを行っており，1 段階目として Twitter データを入力し，生活満足度を推定する複数のモデルを作成し，2 段階目として各モデルの出力を Google Trends データと併せて入力し主観的幸福度を推定した．次にそれぞれのモデルについて説明する．

### ステージ 1. 生活満足度推定モデル

生活満足度の要因ごとに SentenceBERT モデルの構築を行う．学習においては上記 Twitter データセットを用いてテキスト間のコサイン類似度を用いた損失関数を使い距離学習を行い，ポジティブ，ネガティブの状態が同じツイート間の距離が近くなるようなベクトル表現を得る．推論においては，学習データの平均ベクトルと推論対象のテキストのベクトルとのコサイン類似度を求めることで，推論対象のツイートが表す各生活満足度要因の満足状態を-1 から 1 の値として推定する．

### ステージ 2. 主観的幸福度推定モデル

SWB を目的変数とし，説明変数としてステージ 1 のモデルの出力に加えて，マクロ経済指標，Google Trends データを変数ごとに MinMax スケーリングを行って用いた．モデリング手法としては ElasticNet，LightGBM，MLP を用いて比較する．

## 5. 実験

ステージ 1 においてはテストデータとして各国，各年毎に収集したツイートを 100 件ずつ推論し，各要因のスコアの平均値をその国，年の状態とした．ステージ 2 においてはテストデータとして 2019-2021 年の期間，6 カ国における主観的幸福度の推定を行った．評価指標には MSE と $R^2$ を使用した．

ベースラインとしてマクロ経済指標のみのデータを用い，Google Trends データ，Twitter データの有無による比較実験を行った．各実験の共通処理として ElasticNet の回帰係数に基づく特徴量選択を行い，その後 PCA により 5 次元の特徴量化を行った．累積寄与率が 8 割以上となることを確認した．交差検証の方法として時系列交差検証を用いた．

## 6. 結果

各実験条件において共通して高い性能が得られた ElasticNet を用いた場合の実験結果を表 1 に示す．

結果より，マクロ経済指標単独のデータセットよりも Google Trends，Twitter データを併用した場合の方がスコアが良くなることが確認できた．これはマクロ経済指標では取得できなかった国民の状態を SNS データから取得できているためと考えられる．

また Google と Twitter に関しては Twitter 単体の性能が高く，Google と Twitter を併用することで，Google 単体の性能を上回り，Twitter 単体の性能に対しては若干下回るがほぼ同性能となった．

寄与の高かった特徴量の分析に関して，ElasticNet の回帰係数をもとに，Google Trends キーワードからは"結婚式", "離婚", "レジャー", "精神的ストレス", "うつ病","心理的苦痛", "精神的苦痛"の寄与度が高いことが確認できた．Twitter データをもとにした生活満足度の各要因については家庭，教育，生活の要因の寄与が高いことを確認した．

表 1.　評価結果

| | Test | |
|---|---|---|
| | MSE | $R^2$ |
| MFE | 0.0783 | 0.2056 |
| Google | 0.0716 | 0.2740 |
| Twitter | 0.0652 | 0.3387 |
| MFE,Google | 0.0787 | 0.2012 |
| MFE,Twitter | 0.0666 | 0.3242 |
| Google,Twitter | 0.0654 | 0.3366 |
| MFE,Google,Twitter | 0.0671 | 0.3189 |

## 7. まとめ

本研究ではベースラインであるマクロ経済指標に加えて，Google Trends および Twitter データを用いたモデリングを行うことによって主観的幸福度の推定性能が向上することを確認した．その要因として，家庭，教育，生活の状態など，人々の主観的な生活満足度を表現する特徴がデータから得られているためであると特徴量の寄与度の分析の結果から考えられた．

今後の課題として，本検証では対象国を英語圏に絞ったが，対象国を広げること，各国のデータ数を増やすことがあげられる．

### 参考文献

[1] Kjell, O., Giorgi, S., Schwartz, H. A., & Eichstaedt, J. C. (2023). Towards Well-Being Measurement with Social Media Across Space, Time and Cultures: Three Generations of Progress. In World Happiness Report 2023 (11th ed., Chapter 5). Sustainable Development Solutions Network.

[2] Carpi, Tiziana, et al. "A Japanese Subjective Well-Being Indicator Based on Twitter Data." Social Science Japan Journal 25.2 (2022): 273-296.

[3] Algan, Yann, et al. "Well-being through the lens of the internet." PloS one 14.1 (2019): e0209562.

[4] Mairesse, F., Walker, M.A., Mehl, M.R., and Moore, R.K., Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. In Journal of Artificial intelligence Research, 30: 457-500, 2007.

[5] 那須川哲哉, et al. "日本語における筆者の性格推定のための言語的特徴の調査." 言語処理学会第 22 回年次大会発表論文集 (2016): 1181-1184.

[6] Kazuma Mori, Masahiko Haruno, Differential ability of network and natural language information on social media to predict interpersonal and mental health traits, Journal of Personality https://onlinelibrary.wiley.com/doi/abs/10.1111/jopy.12578

[7] Tirotta, Paolo, Akira Yuasa, and Masashi Morita. "Multilevel Sentence Embeddings for Personality Prediction." arXiv preprint arXiv:2305.05748 (2023).