J-025

単語の分散表現を利用した Twitter からの幸福感の抽出に関する考察
Consideration for extraction of the feeling of happiness from tweets of Twitter using distributed representation of word

大沼 智幸† 浅井 紀久夫†
Tomoyuki Onuma Kikuo Asai

1. はじめに

近年、インターネット上に蓄積された大量の情報から何らかの意味をもった知見を抽出し活用したいという要求が顕在化してきている。特に、SNS の個人のつぶやき情報を分析することによって実世界で起きている様々な事象を集合知として解釈し、世の中の動きを把握しようとする取り組みが試みられるようになってきている。

本研究は Twitter のつぶやき情報を収集、分析することで、ユーザが感じている短期的な感情（感情的なうれしさ、楽しさ）を抽出し、これを長期的に観測することで長期的な幸福感（所謂幸福度）に類する指標を推定する方向性を模索している。

本研究では、その初期的な取り組みとして、Twitter のつぶやきから、ユーザが感じている短期的な幸福感を推定することから着手した。

つぶやきからユーザの感情を推定する手法として、教師データのラベル付けの手間を抑えるために、Autoencoder による事前学習の手法を用い、また、ひとつひとつのつぶやきを予め固定長の文書ベクトルに前処理し、Autoencoder の入力とした。文書ベクトルを構成する単語ベクトルは word2vec による単語の分散表現を活用した。

また、今後、短期的幸福感を長期的に観測し考察することに先立ち、短期的幸福感を正しく捉えられているかを考察するために、休日の幸福感が平日よりも大きいという仮説の元、曜日毎の幸福感の比較を試みた。分析対象として 9 日間のつぶやき約 30 万件を無作為に収集し、「Positive／Negative／なし（Neutral）」の三つに分類し、その比率の曜日毎の変化を考察した。

2. 幸福度の定義および計測方法

2.1 幸福度の定義と計測方法

所謂「幸福度」は人々が感じる幸福の度合いを質問紙法などによって調査・考察するものである。代表的なものとしては文献[1]に示される人生満足度（SWLS）が挙げられる。この指標は、「私は自分の人生に満足している」といった五つの質問に対し、その当てはまり具合を 7 段階の選択肢から選び回答するといった方法で導き出される。つまり、幸福度は人生という長い時間軸における感情を計測する指標といえる。

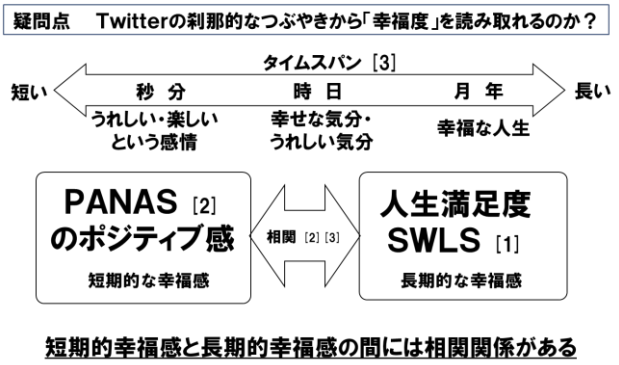

図 1　短期的幸福度と長期的幸福度の関係

2.2 短期的幸福感と長期的幸福感の関係

本研究で分析対象としている Twitter のつぶやきから読み取られる感情は、極めて短期的な時間軸における刹那的な感情（短期的幸福感）と考えられ、それに対し上述の幸福度のような指標は長期的な時間軸における幸福感（長期的幸福感）が反映されていると考えることができる。つまり、短期的幸福感と長期的幸福感は異なる性質のものであり、それらは無関係であるように思われる。

一方で、それらの指標はそれぞれがある程度の関係性（図 1）をもっていることが先行研究により示されている。

文献[2]は、短期的な感情を計測する指標 PANAS におけるポジティブ感情 と上述の「人生満足度」の間に弱い相関があることを明らかにしている。また、文献[3]は『主観的幸福の指標として「感情的幸福」、「幸福度」、「生活満足度」、「ディーナーの人生満足尺度」（順に短期から長期の幸福度）の間にはそれなりの相関がある』とし、いずれの指標も「幸福度」を表す指標とみなすことができると述べている。

これらのことから本研究では、最終的に長期的指標である幸福度との関係を考察するための前段階として、Twitter のつぶやきから短期的な幸福感を抽出する取り組みに着手した。

3. 分析方法

3.1 概要

分析方法の概要を図 2 に示す。本研究では、当初、一般的な手法として n-gram や TF*IDF によるつぶやきのベクトル化、および SVM によるつぶやきの分類を試みた。但し、

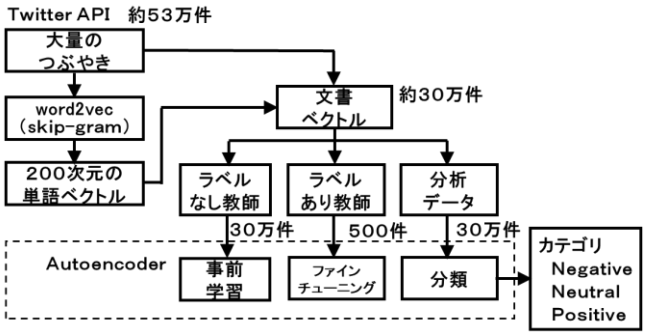

図 2　分析方法

Twitter のつぶやきは一般的に短くて、きれいな日本語ではないことから、ある程度大量のつぶやきを分析する必要性があると考え、そのような事例に適した分析手法を模索した。具体的には文献[4]の Web 新聞記事の分類などを参考に、大量のデータへのラベル付けの手間を抑える目的で Autoencoder による事前学習を活用した。併せて、ひとつひとつのつぶやきを、word2vec による単語の分散表現を活用して文書をベクトル化し、Autoencoder の入力とした。

3.2 データ収集と前処理

Twitter のつぶやきの収集は Twitter API（Streaming API）を用いて行い、Filter 機能により東京周辺の位置情報付きのつぶやきを無作為に収集した。収集期間 9 日間で、トータル約 53 万件のつぶやきを取得した。得られたつぶやきを形態素解析したうえで、そこに含まれる約 700 万語の単語に対し、word2vec（skip-gram）により単語ベクトルを生成した。さらに個々のつぶやきに含まれる単語の単語ベクトルの平均を計算し、各つぶやきの文書ベクトルをそれぞれ生成した。

3.3 分析

本研究の感情分類は、「Positive／Negative／なし（Neutral）」の三分類とし、日毎のそれぞれのつぶやき数を全つぶやきに対する割合で比較することとした。

分析に先立ち、Positive な感情の教師データは「幸せ」「嬉し」「楽し」を含むつぶやきから筆者を含む二名のアノテータの主観により選択し、二人が共通して選択したものを採用した。Negative な教師データも同様に「怖い」「悲しい」「腹立つ」「嫌い」など 12 個の表現を含むつぶやきから、同様に二名のアノテータの主観により選択した。Neutral（Positive でも Negative でもない）については、事前の検索は行わず、任意の 1000 件の文書ベクトルの内容を確認し、明らかに感情を含まないつぶやきを二名のアノテータの主観により選択した。ラベルの付いた教師データの数は約 500 件である。

併せて有効な約 30 万件のラベルなしの文書ベクトルを用いて Autoencoder によって事前学習し、ラベルあり教師データを用いてファインチューニングを行うことで、つぶやきを Positive／Neutral／Negative の三つに分類する分類器を作成した。

最後に約 30 万件のつぶやき（ラベルなし文書ベクトル）を日毎に、分類器によって分類した。

4. 結果と考察

Twitter のつぶやきを曜日毎に分析し、ポジティブ感（幸福感）の割合の変化、特に、平日と休日の感情割合を導出した。休日のポジティブ感が平日よりも高いという直感的な仮説を設け、そのような結果が得られることを期待して分析を行った。

図 3 に 2018 年 1 月 11 日から 19 日の 9 日間に抽出したつぶやきを曜日毎に感情分類した結果を示す。個々のつぶやきの任意の 500 件を二名のアノテータが確認したところ、つぶやきに含まれる感情に基づく感情分類の正答率は 8 割程度であった。

感情比率の曜日毎の変化を捉える試みについては、休日の幸福感の割合の増加を期待したが、結果として、平日と休日（金曜日と土曜日）の結果に有意な差があるとはいえないことがわかった。

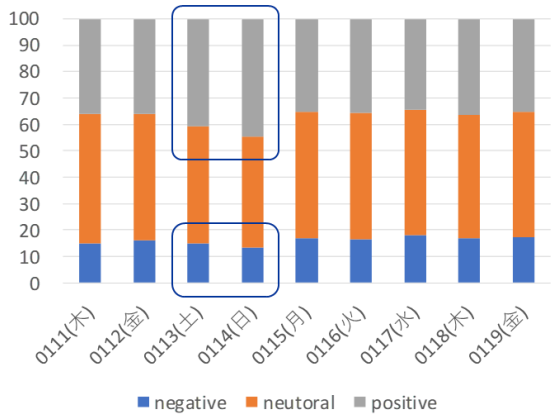

図 3　曜日毎の幸福感の割合の変化

今回の検討の範囲において、平日／休日間の変化傾向を捉えられなかった要因の可能性として①感情分類の精度が十分でないこと、②その週のつぶやきが、たまたま感情比率の変化の少ない週だったという二つが挙げられる。今後、①については前処理の改善や手法の見直しを、②については複数の週に対する分析を実施することを検討したい。

5. おわりに

本研究では文献[4]の Web 新聞記事の分類事例（単語の分散表現と事前学習を利用した分類事例）を参考に、Twitter のつぶやきを大量に分析し、短期的な幸福感を推定するための手順及び考察の流れを模索した。今回の検討の結果として、つぶやきに含まれる感情に基づき、つぶやきを約 8 割の精度で分類することができた。一方で、短期的幸福感の曜日毎の変化については、それを明確に捉えるには至らず課題を残す結果となった。現時点では、充分な精度が得られておらず、また、分析データ数も十分ではない状況である。今後、精度を向上させ、さらに多くのデータ分析を行うことで、短期的感情の時間的変化を捉えられていないという課題の解決を図るとともに、短期的幸福感の分析結果から長期的な所謂「幸福度」を推定することの検討を進める。

謝辞

本研究の一部は JSPS 科研費 17K01052 の助成を受けました。

参考文献

[1] 大石繁宏. 幸せを科学する: 心理学からわかったこと. 新曜社, 2009.
[2] 川人潤子, 大塚泰正, and 甲斐田幸佐. “日本語版 The Positive and Negative Affect Schedule (PANAS) 20 項目の信頼性と妥当性の検討." 広島大学心理学研究 11 (2011): 225-240.
[3] 前野隆司. 幸せのメカニズム 実践・幸福学入門. 講談社, 2013.
[4] 加藤, “教師なしデータを利用した単語の分散表現と事前学習を用いたＷｅｂニュースデータのカテゴリ分類“, 法政大学大学院 修士論文, 2016.

† 放送大学　The Open University of Japan