

うつ病の早期検知に向けたブログデータの時系列分析

Time Series Analysis of Blog Articles for Early Detection of Depression

中野 遼河[†]
Ryoga Nakano安藤 一秋[‡]
Kazuaki Ando

1. はじめに

WHOによると、うつ病は、全世界で成人の5%が苦しんでいると推定される精神障害の一種であり、世界の疾病負担全体に大きく影響を与えていると報告されている[1].

近年では、自然言語処理を精神障害の診断や分析などに応用する研究が注目されている。たとえば、国内におけるうつ病に関する関連研究としては、ブログ記事の内容および感情変化からうつ傾向を検出する研究[2]がある。また、国外の関連研究としては、うつ病になる前に書かれたツイートに注目し、うつ傾向の特徴を抽出し、うつ患者を早期検出する研究[3]がある。

そこで本研究では、ブログの内容からうつ病の兆候が見られた段階で、著者に病院の早期受診・治療を勧める技術の実現を目的とする。本稿では、うつ病と診断される前と後の両方の内容が書かれた記事に注目し、診断前後の記事の特徴を分析する。

2. 研究方針

うつ病は、重度になると自殺につながるおそれもあるため、早期に兆候を発見して治療することが重要である。このような状況の中、関連研究[2]では、うつ傾向があると公言している著者30名のブログと、そうでない著者30名のブログを対象に、両者の特徴を調査し、うつキーフレーズ辞書に基づくうつ検出手法と、感情変動に基づいたうつ検出手法を提案している。うつキーフレーズ辞書に基づく手法については、うつ傾向のある著者のブログ（うつブログ）を69.7%の正解率で、また、そうでない著者のブログ（非うつブログ）を61.1%の正解率で分類できたと述べている。また、感情変動に基づいたうつ検出手法については、最大で86.6%の正解率でうつブログを判定できたが、非うつブログについては65.6%に留まると述べている。

また、関連研究[4]では、英語ツイートを対象として、うつ病を発症する前の1年分のツイートのみを利用して、社会的関与や感情、行動特性などから発症前のうつ病リスクを推定する手法を提案している。

本研究では、ツイートよりも情報量が多いブログを利用する。先行研究[4]と同様、うつ病と診断される前に書かれた記事に、うつ病を早期発見するための特徴が埋もれていると仮定する。また、うつ病と診断された後に書かれた記事にも有用な情報が埋もれている可能性がある。

そこで本研究では、うつ病の診断を受ける前から記事を書いており、うつ病と診断された後も記事を書き続けているブログ著者のデータに注目する。そして、診断日を境に

どのような特徴が埋もれているかを分析した後、うつ病を早期発見する手法を検討する。本研究では、すでに診断前後の記事特徴を分析した[5]。診断前後という区切りだけでなく、時系列で細かく分析することで、早期検知に有効な特徴が見えてくる可能性がある。

本稿では、ブログ記事の中で使用される単語や、投稿数を時系列に基づいて分析する。なお、分析結果をもとに、ブログ著者のうつ病の兆候を早期に発見する手法については、今後検討する。

3. 分析対象データ

本研究では、うつ病の診断を受ける前と、うつ病と診断された後に書かれた記事の両方に注目するため、そのような記事を書いたブログ著者を探さなければならない。本研究では、プロフィール検索を利用でき、利用者数も多いアメーバブログを対象とする。

2024年6月時点で40人のブログ著者が収集できた。次に、ブログ記事をうつ病の診断前1年分の記事と診断後1年分の記事に分けて収集する。収集した結果、一人当たりのブログ件数は、うつ診断前で平均162件、最小7件、最大で1,263件、うつ診断後で平均250件、最小で4件、最大で2,534件となった。診断前後一年という期間を決めて記事を収集しているため、診断前と診断後でも記事数にばらつきが生じる。

4. 時系列分析

4.1 時系列分析を対象者の選出

ブログの内容を時系列で分析するにあたり、著者によって投稿頻度に差があるため、診断前後における最長投稿間隔を調査した。その結果、診断前後において、最長投稿間隔が7日以内の著者は2名、14日以内の著者は6名、21日以内の著者は10名であることがわかった。2名ではデータ数が不十分であるため、集計期間を14日に設定し、診断前後において、最長投稿間隔が14日以内である著者6名を分析対象とした。さらに、この6名中、診断前後において1年間ブログの投稿を継続している著者を3名特定し、この3名（著者A、B、C）を本稿での分析対象とした。残った3名の著者に関するデータの活用方法は今後検討する。

4.2 ブログの時系列分析

3人の著者が書いたブログを時系列分析する。具体的には、投稿数、語句の使用頻度、ネガティブワードの使用頻度などを可視化する。ネガティブワードは、日本語評価極性辞書（用言編）[6]と日本語評価極性辞書（名詞編）[7]でネガティブと判定されているワードを用いる。

本稿では、分析結果のうち、著者ごとに最も特徴的だった結果について述べる。

[†] 香川大学大学院創発科学研究科 Graduate School of Science for Creative Emergence, Kagawa University

[‡] 香川大学創造工学部 Faculty of Engineering and Design, Kagawa University

4.2.1 著者 A のブログ分析結果

うつ病は、強いストレスを長期間受け続けることにより発症するため、ストレスの原因になっている事象の関連語句の使用頻度を時系列で分析する。分析の結果、著者 A で最も特徴的だった点は、ストレス源である「仕事」と「会社」（ストレスワード）の使用頻度と、ネガティブワードの使用頻度との関連性である。著者 A は、まず仕事で適応障害を発症し、2021 年 7 月休職したものの、休職に関する会社とのやり取りでストレスを感じていた。そのため、「仕事」と「会社」それぞれの使用回数を 2 週間ごとの投稿数で割った平均を可視化して関係性を確認した。また、用言ネガティブワードの使用頻度との関連性も確認できたため、ネガティブワードの使用回数を、2 週間ごとの投稿数で割った平均を同時に可視化した。

「仕事」の使用頻度と用言ネガティブワードの使用頻度を図 1 に示す。青と赤のグラフはそれぞれ、用言ネガティブワードの使用頻度と、ストレスワードの使用頻度を示す。図 1 より、「仕事」の使用頻度は休職する前に上昇し、休職した 2021 年 7 月を境に使用頻度が下がっている。ネガティブワードの使用頻度も「仕事」の使用頻度とともに上昇していることがわかる。次に、「会社」の使用頻度と用言ネガティブワードの使用頻度を図 2 に示す。青と赤のグラフはそれぞれ、用言のネガティブワードの使用頻度と「会社」の使用頻度を示す。図 2 より、「会社」は、休職後の 2021 年 7 月を境に使用頻度が上昇し、うつ病と診断された後も使用頻度が上昇し続け、退職した 2021 年 12 月を境に使用頻度が減少している。ネガティブワードの使用頻度も同様のタイミングで上昇し、その後、減少している。これらの結果から、ストレスワードは、ネガティブ度の変動をとらえられる可能性がある。また、ストレスワードの使用頻度は、ネガティブワードの使用頻度と比較して変動が大きいと、機械的に特徴を検知しやすい可能性がある。

4.2.2 著者 B のブログ分析結果

著者 B は、ブログ内で、両親の死が原因で、うつ病を発症したと述べている。著者 B の特徴は、投稿数と名詞ネガティブワードの使用頻度の関係である。うつ病による体調の変化が投稿数に反映されるのではないかと考え、投稿数を 2 週間ごとに可視化した。その結果、著者 B もネガティブワードの使用頻度との関連性が確認できたため、同時に可視化した。投稿数と名詞ネガティブワードの使用頻度を図 3 に示す。青と赤のグラフはそれぞれ名詞のネガティブワードの使用頻度と、投稿数を表している。図 3 より、うつ病と診断された、2019 年 10 月前後に投稿数が減少していることがわかる。投稿数の減少とは対照的に、名詞のネガティブワードの使用頻度は上昇している。うつ病によって気分がネガティブになり、投稿数が減少したと考えられる。このことから、投稿数の変化によって、うつ病による体調の変化を捉えられる可能性がある。

4.2.3 著者 C のブログ分析結果

著者 C は、ブログ内で、父親の介護が原因でうつ病を発症したと述べている。著者 C は著者 A とは対照的に、ストレスに関連するワードの使用頻度の変化が診断前後で変化しなかった。このような著者は、慢性的なストレスに気づいていない可能性がある。例えば、仕事や家事など、日常生活の中で常にストレスを感じている場合、個々のストレスイベントに対する意識が薄くなり、関連するワードの使

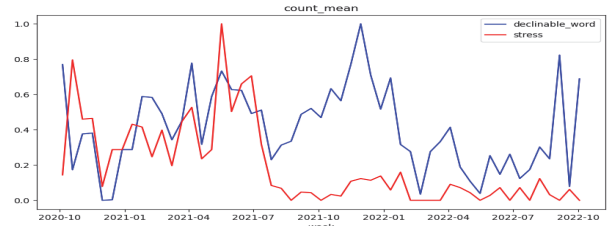


図 1 「仕事」と用言ネガティブワードの使用頻度

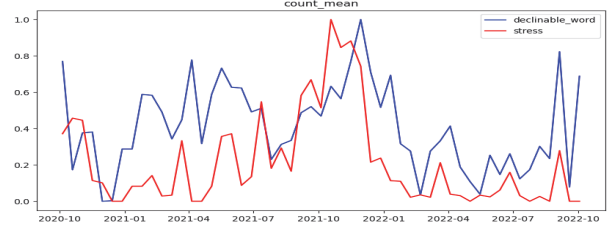


図 2 「会社」と用言ネガティブワードの使用頻度

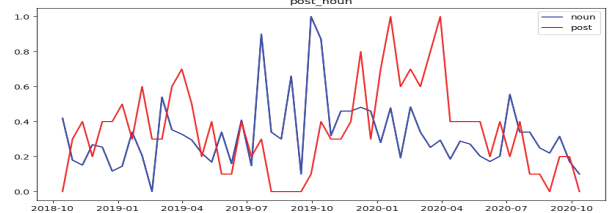


図 3 投稿数とネガティブワードの使用頻度

用頻度にも変化が現れないと考えられる。そのため、ストレスに気づきにくい著者に対して、うつ病を検知する仕組みを構築する必要がある。

5. おわりに

本稿では、うつ病と診断される前後の記事を書いたブログ著者に注目し、特定の著者に対して、ブログ記事の中で使用される単語や、投稿数を時系列分析した。分析結果より、ストレス源の特定と、それに関連する語句の使用頻度の増加、投稿数の変化が、うつ病の早期検知に役立つ可能性があることがわかった。また、ストレスに気づきにくい著者の存在が確認できたため、このような著者に対してうつ病を検知する仕組みの構築が課題となる。

今後は、データ分析で得られた知見を用いて、ブログ著者のうつ病の兆候を早期に発見する手法について検討する。

参考文献

- [1] WHO, "Depression", <https://www.who.int/en/news-room/fact-sheets/detail/depression>, 2021.
- [2] 松本他, "うつキーフレーズと感情変動に基づくブログからのうつ検出手法", NLP2012 論文集, pp.1126-1129, 2012.
- [3] 高須他, "大規模ツイートデータを用いたメンタルヘルス不調者の推測", 2022 年度人工知能学会全国大会 (第 36 回) 論文集, 4 pages, 2022.
- [4] M. De. Choudhury, et al., "Predicting Depression via Social Media", Proc. of the Seventh International AAAI Conference on Weblogs and Social Media, pp.128-137, 2013.
- [5] 中野他, "うつ病の早期検知に向けたブログデータの分析", 第 22 回情報科学技術フォーラム講演論文集, 第 2 分冊, pp.377-378, 2023.
- [6] 小林他, "意見抽出のための評価表現の収集", 自然言語処理, Vol.12, No.3, pp.203-222, 2005.
- [7] 東山他, "述語の選択選好性に着目した名詞評価極性の獲得", NLP2008 論文集, pp.584-587, 2008.