6W-02

大規模言語モデルを用いたSNS投稿からの精神疾患の推測

滝波秋穂† 岸本泰士郎†† 狩野芳伸†††
静岡大学† 慶応大学†† 静岡大学†††

Abstract

SNSに投稿された文章を用いて発信者に精神疾患の傾向があるかどうか、専門医の診断付き患者のツイートデータから、大規模言語モデルを用いて推測を試みる。昨今精神疾患の患者数増加が社会的な問題となっているが、SNSでの日々の発信を用いることで、精神疾患の早期発見に資する可能性がある。SiSP状況別極性単語辞書ベースのSVMと大規模言語モデル（GPT-4）それぞれのアプローチで疾患有無の推測を行い、結果を比較した。SVMのほうがGPT-4より高い総合性能を示し、極性単語が相当程度有効であると同時に、プロンプト作成の難しさ、表層的な特徴のみを捉えていると考えられることからも、現在の大規模言語モデルの直接的な利用は難しいことを示した。

1. はじめに

昨今精神疾患の患者数増加が社会的な問題となっている。精神疾患の診断には迅速な診断が難しい、重症化するまで受診がされないといった問題があり、適切な治療の遅れは病状の悪化を招き、復帰を難しくしてしまう。本研究では、早期発見に資する予備診断的な利用を想定し、日々のSNSの投稿を分析し推測を試みる。

深谷ら[1]は、抗うつ剤に関するツイートをしたSNSユーザーを抽出し、SVMによりそうでないユーザーと分類したが、事前に抽出した単語のみしか分析できず、文脈も考慮されていない。Murarkaら[2]は、アメリカの投稿型ソーシャルサイトRedditへの投稿を、機械学習を用いて5つの精神疾患に分類した。高須ら[3]は、Twitter上で薬剤に関する言及をしかつプロフィールに疾患・症状の訴えに関する単語を含むユーザーについて、そうでないユーザーと深層学習を用いて分類を行ったが、抽出したアカウントには正確な診断情報がなく、アカウントに偏りがある可能性がある。Kailaiら[4]は大規模言語モデルによりメンタルヘルス分析を行うMentaLLaMAを開発したが、投稿は英語を想定しており、日本語のモデルはまだない。

本研究では日本語を対象に、診断付きのツイートアカウント群を用いて、大規模言語モデル（以下、LLM）と単語ベースの手法それぞれで疾患有無の推測を試みた。

2. データ

UNDERPINプロジェクト[5]で収集した精神科医の診断付きTwitterアカウントのツイートを用いる。本研究で使用するアカウントは、健常者23人、うつ病17人、双極性障害19人である。本研究ではアカウント本人のテキストを対象とし、ツイート中の被験者発話以外、すなわち絵文字やリンク、リツイートを除外した。LLMとして用いたGPT-4（gpt-4-1106-preview）の最大入力長を考慮して、同一アカウントのツイート群を、各ツイートが分割されないように4000文字以内で結合したファイルを作成した。

3. 実験

ツイートを入力として、紐づけられた診断を正解とし、その推測をGPT-4とSVMによって行い比較した。アカウント当たりのツイート数は偏りがあるため、ファイル単位での訓練評価と、アカウント単位での訓練評価（SVMは全投稿、GPT-4は先頭のファイル一つ）も行った。

GPT-4のプロンプトは「この投稿主のうつ病及び双極性障害の有無を推測してください。」と尋ね、疾患の特徴をいくつか説明したうえで、ツイート事例を健常者：有病者＝1:3でfew shot promptingした。1:1では健常という答えばかりになってしまったためである。

Inferring mental illness from SNS posts by large language model
†Takinami Akiho, Shizuoka University
††Kishimoto Taishiro, Keio University
†††Kano Yoshinobu, Shizuoka University

SVM の学習特徴量には高田ら[6]の作成した状況別感情極性日本語辞書 SiSP を利用した。SiSP では 25,520 単語について、20 の異なる状況に対しクラウドワーカーによって 10 件ずつ極性の投票が行われた。アカウントごとの辞書内単語出現数に各状況別極性の投票重みを掛け算した。これを特徴量として SVM の学習を行い、件数が非常に少ないために。ハイパーパラメータのグリッドサーチの訓練-検証を 10 分割交差検証、その評価をさらに 10 分割交差検証し、各評価尺度の平均値を最終結果とした。

4. 結果

状況別極性の重みは、健常・有病に関わらず、ポジティブな単語は状況＝コミュニケーションの影響が、ネガティブな単語では状況＝経済の影響が大きかった。そこでアカウントごとのツイートについて、コミュニケーション及び経済の極性重みを算出し、それらを特徴量として SVM での推測を試みた。

SVM 及び LLM それぞれについて、アカウント単位とファイル単位それぞれで訓練・評価した（表 1）。ファイル単位のラベル比率は SVM では健常 859:有病 1467 で、LLM では 1:1 になるようランダムに 100 件抽出した。総合性能としては SVM の方が LLM よりも高かった。

	正解率	適合率	再現率	F1 値
SVM-A	0.740	0.559	0.608	0.570
SVM-F	0.763	0.745	0.735	0.735
LLM-A	0.559	0.917	0.305	0.458
LLM-F	0.535	0.750	0.118	0.203

表 1. 推測結果(-A アカウント -F ファイル 単位)

5. 考察

SVM ではアカウント単位でもファイル単位でも同程度の正解率だったが、F1 値の値には大きく差がある。アカウント単位では交差検定のうち推測が有病または健常に偏る事例があり、訓練データ不足で学習がうまくいかなった可能性がある。また、アカウント単位でもファイル単位でも、推測に失敗したアカウントには共通の特徴が見られた。また、健常者より有病者の方が極性に関わる単語の出現が多い、つまり感情表現の多い投稿をする傾向にあった。SVM で健常と誤分類されたアカウントの投稿は、日常を淡々と綴ったものやニュースの感想など、感情表現が少なかった一方、有病と誤分類された投稿では、感情豊かな表現が見られた。これは人による SNS 表現のスタイルで、単語の極性のみで推測を行うのは限界があることが分かる。

LLM の結果では、正解率が低いうえ、適合率に偏り健常の出力例が多い。プロンプトをさまざま試行錯誤した過程ではどちらかに偏ることが多く、比較的バランスが良いのがこのプロンプトであった。有病の正解事例には、ネガティブな発言や体調不良に関する発言が非常に多かったが、そうした特徴がない投稿は推測ができず、LLM は目視の判断に近い結果とも言える。プロンプト作成の難しさ、表層的な特徴のみを捉えていると考えられることからも、現在の LLM の直接的な利用は難しいと考えらえる。

6. 参考文献

[1]深谷拓吾, 川西直, 長谷川晃朗,田近安蘭,小川雄右,堀越勝,古川壽亮,武内良男.”うつ傾向推定に向けた抗うつ剤服用の投稿を起点として Twitter 解析の初期検討”言語処理学会第 21 回年次大会発表論文集(2015)

[2]AnkitMurarka,BalajiRadhakrishnan,SushmaRavichandran.”Classification of mental illnesses on social media using RoBERTa.”In Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis,pages59-68,online.Association for Computational Linguistic.(2020)

[3]高須遼, 中村啓信, 岸本奉士郎, 狩野芳伸.”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論文集(2022)

[4] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Sophia Ananiadou, Jimin Huang.” MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models.” arXiv:2309.13567 (2023)

[5]Kishimoto, Taishiro, et al. ″Understanding psychiatric illness through natural language processing (UNDERPIN): Rationale, design, and methodology.″ Frontiers in Psychiatry 13 (2022): 954703.

[6]高田篤志,狩野芳伸,山崎俊彦.”状況別感情日本語辞書の作成とその活用”言語処理学会第 28 回年次大会発表論文集(2022)