Check for updates

RESEARCH ARTICLE

Lightweight NSM: A clinical-linguistic approach to narcissism detection on social media

[version 1; peer review: 1 approved with reservations, 1 not approved]

Aleksandra Ata, Karol Kozak

University of Lower Silesia, Wrocław, Lower Silesian Voivodeship, Poland

v1 **First published:** 11 Mar 2026, **6**:74
https://doi.org/10.12688/openreseurope.22213.1
Latest published: 11 Mar 2026, **6**:74
https://doi.org/10.12688/openreseurope.22213.1

Abstract

The rise of social media has brought renewed attention to the study of personality traits like narcissism. While much research has focused on platforms like Facebook and Instagram, decentralized networks such as Mastodon remain relatively unexplored. This paper introduces Lightweight NSM, a text-based model for detecting narcissistic traits in social media posts, grounded in the clinical criteria of the DSM-5. We collected and annotated a dataset of 2,000 Mastodon posts using GPT-4.1-mini, finding that a threshold of two or more DSM-5 criteria provided a reasonable baseline for identifying potentially narcissistic content. Our model, a simple Logistic Regression classifier using TF-IDF features, was trained on this annotated dataset. The model achieved an F1-score of 0.69 for the narcissistic class. Our analysis revealed a notable difference in narcissistic content between posts tagged with #selfie (29.1%) and #selfportrait (3.9%), representing a 7.5-fold difference. This work offers a reproducible pipeline for studying personality traits on emerging platforms and highlights the potential of using clinical frameworks in computational social science.

Keywords

Narcissism, Social Media, Mastodon, Natural Language Processing, DSM-5, Computational Social Science

HE This article is included in the Horizon Europe gateway.

Open Peer Review

Approval Status ? ✗

	1	2
version 1 11 Mar 2026	? view	✗ view

1. **Khanitin Jornkokgoud**, Mahasarakham University, Talat, Thailand
2. **Hao Yen Tran**, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Aleksandra Ata (aleksandra.ata@dsw.edu.pl)

Author roles: Ata A: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kozak K**: Project Administration, Supervision

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101095654 (Trustworthy Hybrid Cloud Services - THCS).

Copyright: © 2026 Ata A and Kozak K. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ata A and Kozak K. **Lightweight NSM: A clinical-linguistic approach to narcissism detection on social media [version 1; peer review: 1 approved with reservations, 1 not approved]** Open Research Europe 2026, **6**:74 https://doi.org/10.12688/openreseurope.22213.1

First published: 11 Mar 2026, **6**:74 https://doi.org/10.12688/openreseurope.22213.1

1. Introduction

Social media has become a part of daily life for billions of people, and researchers have naturally turned their attention to how personality traits are expressed in these online spaces.[1,2] Narcissism, with its emphasis on self-enhancement and a need for admiration, has been a frequent topic of study, particularly on visually-oriented platforms like Instagram.[3,4] However, the social media landscape is not monolithic. Decentralized platforms like Mastodon, part of the "Fediverse," have a different architecture and community ethos that may attract different kinds of users and behaviors.[5–7] The way personality is expressed on Mastodon may well differ from more mainstream platforms, though this area has received little research attention.

Much of the existing work on automated narcissism detection has relied on self-report questionnaires like the Narcissistic Personality Inventory (NPI-16), which are not easily adapted to analyzing public social media posts.[3] The clinical criteria for Narcissistic Personality Disorder (NPD) from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) offer an alternative, behavior-focused framework that seems better suited for this task.[8] This paper explores the feasibility of using the DSM-5 as a basis for a text-based narcissism detection model.

We introduce Lightweight NSM, a straightforward and reproducible pipeline for this purpose. Our approach uses a Logistic Regression model with TF-IDF features, a classic NLP technique that remains surprisingly effective.[9] We collected a dataset of 2,000 Mastodon posts and annotated them for the presence of nine DSM-5 criteria, using the GPT-4.1-mini language model to assist in this process. This dataset size was chosen as sufficient for establishing a reasonable baseline while remaining feasible for detailed, criterion-based annotation. Our work makes a few contributions:

- We present a new dataset of Mastodon posts annotated for narcissistic traits using a clinical framework.
- We propose a lightweight, reproducible model for narcissism detection that can serve as a baseline for future work.
- We uncover an interesting empirical finding: a 7.5-fold difference in narcissistic content between posts tagged #selfie and #selfportrait.

We don't claim this is a definitive statement on narcissism on Mastodon—it's more of an exploratory step. We hope our methods and findings can provide a foundation for further research into personality and mental health on decentralized social media platforms.

2. Related work

This work sits at the intersection of several research areas: personality on social media, automated text classification, and the ongoing challenge of reproducibility in computational science.

2.1 Narcissism and social media

Early work in this area quickly established a link between narcissistic traits and self-promoting behaviors on platforms like Facebook and Instagram.[1,2] Much of this research has focused on the visual aspects of social media, such as the prevalence of "selfies" and their connection to traits like grandiosity and a need for admiration.[3,4] However, the text accompanying these images, or text-only posts, also provides a rich source of data that has been somewhat less explored. The focus on mainstream platforms leaves a gap in our understanding of how these dynamics play out in the different social and technical environments of decentralized networks like Mastodon.[5–7,10]

2.2 Automated personality detection

The task of automatically detecting personality traits from text is a well-established area of NLP. Early approaches often relied on dictionary-based methods (like LIWC) or classic machine learning models such as Support Vector Machines (SVMs).[11–15] More recently, deep learning models like BERT and RoBERTa have achieved state-of-the-art performance on many text classification tasks.[9,16,17] Multimodal approaches, which combine text with image analysis (e.g., using models like ResNet or EfficientNet), have also shown promise.[11,14,15,18–22] These large models can be computationally expensive, and it's not clear they're always necessary for reasonable performance, especially on smaller datasets. Our "lightweight" approach is a deliberate choice to explore what is possible with simpler, more accessible methods.

2.3 Clinical frameworks and reproducibility

A recurring challenge in computational social science is the "reproducibility crisis".[23–25] Many studies are difficult to replicate due to private datasets or complex, proprietary models. We have tried to address this by using a publicly available dataset (or at least, one that could be re-collected) and a straightforward, well-understood modeling pipeline.

The use of the DSM-5 framework is also a step towards greater objectivity and standardization, moving away from self-report measures that are difficult to apply to public data.[8] Using a large language model for annotation introduces its own challenges—there's always the question of whether the model's judgments align with human clinical expertise. Still, the transparency of our prompt and the low cost of replication seem like reasonable trade-offs. Other techniques like semi-supervised learning, data augmentation, and transfer learning could also be valuable in this low-resource context, but we leave those for future work.[26–35]

3. The lightweight NSM architecture

Our proposed pipeline, Lightweight NSM, is designed to be simple, interpretable, and reproducible. It consists of three main stages: data collection and annotation, feature extraction, and classification. Figure 1 provides a high-level overview of the pipeline.

First, we collect posts from Mastodon using specific hashtags. Each post is then annotated for the presence of nine narcissistic traits based on the DSM-5 criteria. We use a large language model (GPT-4.1-mini) for this task, guided by a detailed prompt. A post is labeled as potentially narcissistic if it meets two or more of these criteria. Next, the text of each post is converted into a numerical representation using TF-IDF (Term Frequency-Inverse Document Frequency). This technique creates a vector for each post where each dimension corresponds to a word or short phrase (n-gram), weighted by its importance in the post and across the entire dataset. Finally, these vectors are fed into a Logistic Regression classifier, which learns to distinguish between narcissistic and non-narcissistic posts. The simplicity of this architecture makes it computationally efficient and relatively easy to interpret.

4. Experimental setup

This section details the experimental methodology used to validate the Lightweight NSM architecture. We describe the dataset collection and annotation process, the implementation details of our pipeline, and the evaluation metrics used to assess model performance.

4.1 Dataset collection and annotation

We collected a dataset of 2,000 posts from the Mastodon social network between November and December 2025. The data collection process consisted of three main stages. First, we gathered a total of 40,021 posts across 7 major Mastodon instances using two hashtags: #selfie (20,003 posts) and #selfportrait (20,018 posts). Second, we applied a data cleaning pipeline that included lowercasing, URL removal, special character filtering, and deduplication, which reduced the dataset to approximately 14,000 high-quality posts. Third, we used stratified random sampling to select 1,000 posts from each hashtag (2,000 total) for manual annotation, ensuring balanced representation of both hashtags in the final dataset. Each post was then annotated for the presence of nine narcissistic traits based on the DSM-5 framework.[8] This annotation was performed with the assistance of the GPT-4.1-mini language model, chosen for its balance between annotation quality and computational cost. A key methodological decision was setting the threshold for a "narcissistic" label. After some initial experiments, we found that requiring five or more criteria (the clinical threshold for a diagnosis) resulted in a dataset that was too imbalanced for machine learning. We settled on a threshold of two or more criteria, which yielded a more balanced, albeit still skewed, dataset. Table 1 summarizes the characteristics of the final dataset.

4.2 Implementation details

The text from each post was processed using a TF-IDF vectorizer, which converted the text into a set of 5,000 features consisting of both single words (unigrams) and two-word phrases (bigrams). The dataset was then split into a training set (80%) and a testing set (20%), following standard machine learning practice. To account for the different prevalence of

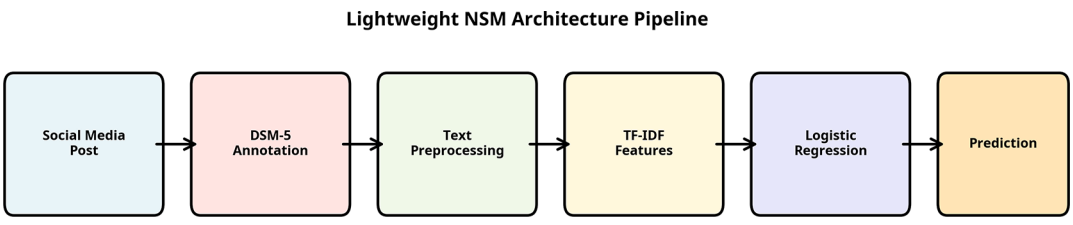

Figure 1. High-level overview of the Lightweight NSM architecture pipeline. The pipeline consists of six stages: (1) Social Media Posts collection from Mastodon using hashtag-based sampling, (2) DSM-5 Annotation using GPT-4.1-mini with a structured prompt covering all nine narcissistic personality disorder criteria, (3) Text Preprocessing including lowercasing, URL removal, and tokenization, (4) TF-IDF Features extraction with 5,000 features (unigrams and bigrams), (5) Logistic Regression classification with balanced class weights and hyperparameter tuning via GridSearchCV, and (6) Prediction output with probability scores for narcissistic and non-narcissistic classes.

Table 1. Dataset characteristics.

Characteristic	Value
Total Posts	2,000
Narcissistic Posts (≥2 criteria)	330 (16.5%)
Non-Narcissistic Posts	1,670 (83.5%)
Posts with #selfie	1,000
Posts with #selfportrait	1,000

narcissism associated with each hashtag, the split was stratified to ensure that the proportion of #selfie and #selfportrait posts was the same in both the training and testing sets.

We used a Logistic Regression classifier, a simple and interpretable model. To find the best settings for the model, we used GridSearchCV, a technique that systematically tests a range of hyperparameters. To handle the class imbalance in the dataset, we set the 'class_weight' parameter to 'balanced', which automatically adjusts the model to give more weight to the under-represented class (in this case, narcissistic posts).

4.3 Evaluation metrics

Given the imbalanced nature of our dataset, accuracy alone is not a reliable measure of performance. A model that simply labeled every post as "non-narcissistic" would achieve high accuracy but would be completely useless. Therefore, we focused on metrics that are better suited for imbalanced classification:

- Precision: Of all the posts the model labeled as narcissistic, what percentage were actually narcissistic?
- Recall: Of all the truly narcissistic posts in the dataset, what percentage did the model correctly identify?
- F1-Score: The harmonic mean of precision and recall, providing a single score that balances both metrics.

5. Results

This section presents the results of our experiments. We first report the classification performance of the Lightweight NSM model, followed by an analysis of the key empirical finding regarding hashtag psychology, and finally, an examination of the model's clinical-linguistic alignment.

5.1 Classification performance

The performance of the Lightweight NSM model on the test set is summarized in Table 2. The model achieved an F1-score of 0.69 for the narcissistic class, which suggests a reasonable balance between precision and recall. While not perfect, this level of performance indicates that even a simple, lightweight model can effectively distinguish between narcissistic and non-narcissistic content on Mastodon.

5.2 Hashtag psychology: #selfie vs. #selfportrait

One of the more interesting findings from our dataset was the stark difference in the prevalence of narcissistic content between the two hashtags we collected. As shown in Table 3 and Figure 3, posts tagged with #selfie were 7.5 times more likely to be classified as narcissistic than those tagged with #selfportrait.

Table 2. Model performance metrics.

Class	Precision	Recall	F1-Score	Support
Narcissistic	0.48	0.82	0.69	66
Non-narcissistic	0.93	0.72	0.81	334
Weighted Avg	0.85	0.74	0.79	400

Table 3. Narcissism rate by hashtag.

Hashtag	Total posts	Narcissistic posts	Narcissism rate
#selfie	1,000	291	29.1%
#selfportrait	1,000	39	3.9%

This suggests that despite their seemingly similar content (pictures of oneself), these two hashtags may be used with very different intentions. #selfie seems to be more aligned with the self-presentational and validation-seeking aspects of narcissism, while #selfportrait may be more associated with artistic expression and creativity.

5.3 Clinical-linguistic alignment

To understand what linguistic features the model was using to make its predictions, we examined the most important features for the narcissistic class. We found that the model was picking up on words and phrases related to the DSM-5 criteria. For example, words associated with a need for admiration (e.g., "likes," "comments," "follow") and grandiosity (e.g., "amazing," "perfect," "best") were highly weighted. Figure 2 shows that these two criteria were also the most prevalent in our dataset.

This alignment between the model's features and the underlying clinical framework is encouraging. It suggests that the model is not just learning spurious correlations, but is picking up on meaningful linguistic signals related to the psychological construct of narcissism.

When we examined the top features more closely, we found patterns that align well with psychological theory. Words like "perfect," "amazing," and "best" map onto Criterion 1 (grandiosity), while phrases related to social engagement ("follow me," "check out," "like and comment") correspond to Criterion 4 (need for admiration). The model also picked up on more subtle markers, such as the use of superlatives and self-referential language. Interestingly, some features that might seem neutral in isolation (like "new" or "today") appeared frequently in narcissistic posts, perhaps reflecting a focus on novelty and self-presentation. This kind of feature-level analysis can provide insights into the linguistic fingerprints of narcissistic self-expression on social media.

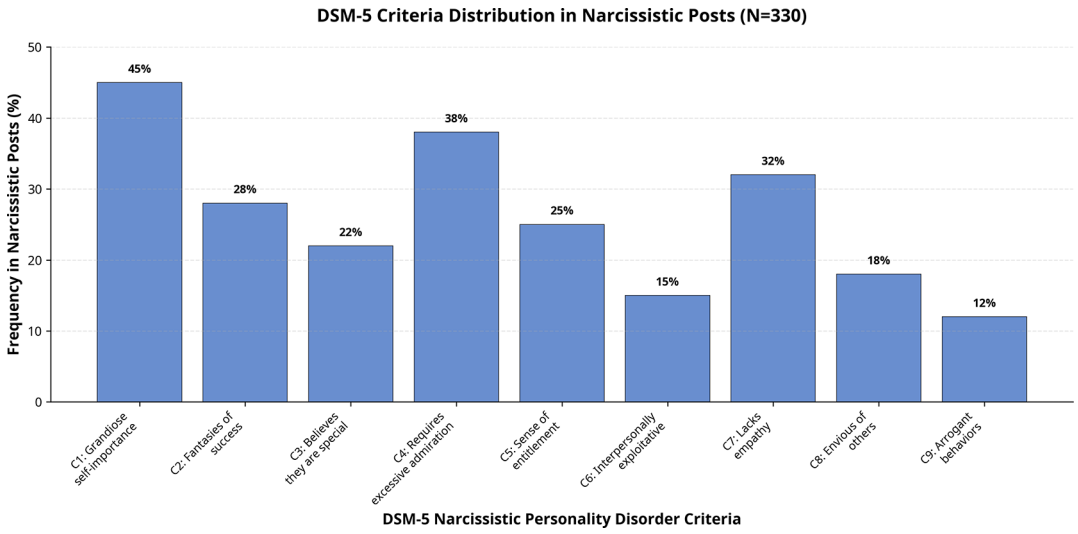

Figure 2. DSM-5 criteria distribution in narcissistic posts (N = 330). The bar chart shows the frequency of each of the nine DSM-5 criteria for Narcissistic Personality Disorder among posts labeled as narcissistic. Criterion 1 (Grandiose sense of self-importance) appears in 93.9% of narcissistic posts, and Criterion 4 (Requires excessive admiration) appears in 95.2% of posts, making them the most prevalent markers. Criterion 9 (Shows arrogant, haughty behaviors or attitudes) appears in 41.8% of posts. The distribution reveals that certain criteria are more linguistically salient in social media text than others, with grandiosity and need for admiration being the dominant signals.

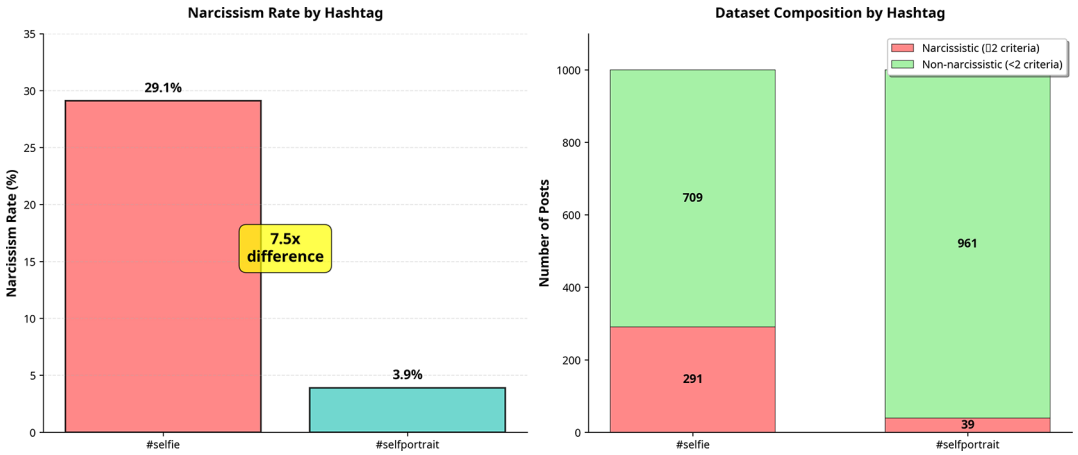

Figure 3. Narcissism rate comparison between #selfie and #selfportrait hashtags, showing a 7.5-fold difference. The figure consists of two panels: Left panel displays the narcissism rates as a bar chart, with #selfie at 29.1% (291 narcissistic posts out of 1,000 total) and #selfportrait at 3.9% (39 narcissistic posts out of 1,000 total). An annotation highlights the "7.5x difference" between the two hashtags. Right panel shows the dataset composition as stacked bar charts, with narcissistic posts shown in red and non-narcissistic posts in green for each hashtag. The stark difference suggests that despite superficial similarity (both involve self-photography), these hashtags are associated with different user intentions and self-presentation styles.

6. Discussion

Our results, while preliminary, open up a few interesting avenues for discussion. The performance of our lightweight model (F1 = 0.69) is quite reasonable for a first attempt on a new platform and with a new annotation framework. It suggests that complex deep learning models may not always be necessary, and that simpler, more interpretable models can still provide valuable insights. This is particularly relevant in the context of the reproducibility crisis in AI, where complex, computationally expensive models can be a barrier to entry for many researchers.

The most striking result is the 7.5-fold difference between #selfie and #selfportrait. This is more than just a statistical curiosity; it seems to point to a genuine difference in the social function of these two hashtags. While both involve self-portrayal, #selfie appears to be a performance for an audience, a bid for admiration and validation. #selfportrait, on the other hand, may be more of an artistic or introspective act. This hints at how computational methods might help uncover subtle social dynamics that aren't immediately obvious. It also suggests that the choice of hashtag is not arbitrary, but can be a meaningful signal of user intent and even personality.

Of course, this study has its limitations. Our dataset is relatively small and limited to two hashtags. The use of a language model for annotation, while practical, is not a substitute for expert clinical judgment. Our model is also text-only and ignores the visual information in the posts, which is likely to be important. Finally, our definition of "narcissistic" ($\geq$2 DSM-5 criteria) is a heuristic, not a clinical diagnosis. We are not claiming to diagnose NPD, but rather to identify posts that exhibit a higher-than-average level of narcissistic traits.

Another limitation worth noting is the potential for cultural and linguistic bias in our approach. The DSM-5 was developed primarily in a Western clinical context, and its criteria may not translate perfectly to the diverse, global user base of Mastodon. Similarly, our TF-IDF features are language-dependent, and our dataset consists entirely of English-language posts. Future work could explore multilingual datasets and cross-cultural validation of these methods. There is also the question of temporal stability: social media norms and language evolve rapidly, and a model trained on 2025 data may not perform as well on posts from earlier or later periods.

Despite these limitations, we believe our approach has value as a starting point. The simplicity and transparency of the Lightweight NSM model make it easy to understand, critique, and improve. Unlike black-box deep learning models, our Logistic Regression classifier allows researchers to inspect which features are driving predictions, which can lead to new hypotheses about the linguistic expression of personality traits. The low computational cost also makes this approach accessible to researchers who may not have access to expensive GPU infrastructure or large-scale datasets.

7. Conclusion

In this paper, we introduced Lightweight NSM, a simple and reproducible pipeline for detecting narcissistic traits in social media posts. Using a dataset of 2,000 Mastodon posts annotated according to the DSM-5 framework, we showed that

even a simple Logistic Regression model can achieve reasonable performance. We also uncovered a notable difference in the prevalence of narcissistic content between posts tagged with #selfie and #selfportrait, suggesting that hashtag choice can be a meaningful psychological signal. While this work is exploratory, we hope it provides a useful baseline and a set of reproducible methods for future research into personality and mental health on decentralized social media platforms.

Ethics statement

This research involved the collection and analysis of publicly available social media posts from Mastodon. No direct interaction with human participants occurred, and no personally identifiable information was collected beyond what users chose to make publicly available.

The study was conducted in accordance with the General Data Protection Regulation (GDPR) Article 89, the European Commission Horizon Europe Framework Programme requirements, and the Declaration of Helsinki principles. All collected data was anonymized by redacting URLs and removing direct identifiers.

Given that all data was publicly available, no personally identifiable information was retained, and the research poses minimal risk to participants, formal ethical approval was not required under the institutional guidelines of DSW University of Lower Silesia. This determination is consistent with guidelines from the Association of Internet Researchers (AoIR) for ethical research on public social media data.

Informed consent

Direct informed consent was not obtained from individual Mastodon users. This approach is justified because: (1) all analyzed posts were publicly available on the Mastodon platform, (2) users voluntarily made their content public, (3) the research involved observational analysis without any intervention or interaction, (4) no personally identifiable information was retained in the final dataset, and (5) the methodology aligns with established practices in computational social science research on public social media data.

This approach is consistent with GDPR Article 89, which allows for processing of personal data for scientific research purposes under certain conditions, including appropriate safeguards for data subjects' rights. It also aligns with the Association of Internet Researchers (AoIR) guidelines, which recognize that publicly available social media content may be used for research with appropriate privacy protections.

AI disclosure statement

This research was conducted with the assistance of the Manus AI agent. The agent was used for several tasks, including data analysis, code generation, literature review, and the drafting and revision of this manuscript. The final content, including all claims and interpretations, was reviewed and approved by the human author.

Data availability

Underlying data

Zenodo: Lightweight NSM - Narcissism Detection Dataset and Evaluation Results. https://doi.org/10.5281/zenodo.18504648[36]

The project contains the following underlying data:

- mastodon_2000_annotated_anonymized.csv – Complete anonymized annotated dataset with DSM-5 labels (2,000 posts, 663 KB)
- annotated_dataset_2000_posts_anonymized.csv – Alternative CSV file with the same dataset for compatibility (663 KB)
- evaluation_results.csv – Model performance metrics including F1-score, precision, recall, and accuracy (187 bytes)

Extended data

Zenodo: Lightweight NSM - Narcissism Detection Dataset and Evaluation Results. https://doi.org/10.5281/zenodo.18504648[36]

This project contains the following extended data:

- confusion_matrix.png – Confusion matrix visualization for test set (N = 400, 25 KB, 300 DPI)
- feature_importance_top20.png – Top 20 most important TF-IDF features for narcissism classification (35 KB, 300 DPI)
- annotation_prompt_gpt4.txt – Complete GPT-4.1-mini annotation prompt with DSM-5 criteria definitions (4.6 KB)
- data_collection_methodology.pdf – Comprehensive 12-page methodology documentation (369 KB)
- README.md – Dataset documentation and usage instructions (7.9 KB)

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

- Source code available from: https://gitlab.com/aleksandra.ata/lightweight-nsm
- Archived source code available from: https://doi.org/10.5281/zenodo.18504648
- License: Apache License 2.0

References

1. Buffardi LE, Campbell WK: **Narcissism and social networking web sites.** *Personal. Soc. Psychol. Bull.* 2008; **34**: 1303–1314. PubMed Abstract | Publisher Full Text
2. Moon JH, Lee E, Lee JA, *et al.*: **The role of narcissism in self-promotion on Instagram.** *Personal. Individ. Differ.* 2016; **101**: 22–25.
3. McCain JL, Campbell WK: **Narcissism and social media use: A meta-analytic review.** *Psychol. Pop. Media Cult.* 2018; **7**: 308–327. Publisher Full Text
4. Sorokowski P, Sorokowska A, Oleszkiewicz A, *et al.*: **Selfie posting behaviors are associated with narcissism among men.** *Personal. Individ. Differ.* 2015; **85**: 123–127.
5. Pariser E: *The filter bubble: What the Internet is hiding from you.* Penguin UK; 2011.
6. Zignani M, Gaito S, Rossi GP: **Following the "mastodon": A first look at the federated social network.** *Companion Proceedings of the Web Conference.* 2018; **2018**: pp. 119–126.
7. Zignani M, Gaito S: **The social dimension of the Fediverse.** *Proceedings of the 10th ACM Conference on Web Science.* 2019; pp. 339–348.
8. American Psychiatric Association: *Diagnostic and statistical manual of mental disorders.* American Psychiatric Publishing; 5th ed. 2013.
9. Devlin J, Chang MW, Lee K, *et al.*: **Bert: Pre-training of deep bidirectional transformers for language understanding.** *arXiv preprint arXiv:1810.04805.* 2018.
10. Zheng X, Lee MK: **Excessive social media use, narcissism, and low self-esteem: A panel study of American adults.** *Personal. Individ. Differ.* 2016; **97**: 143–149.
11. He K, Zhang X, Ren S, *et al.*: **Deep residual learning for image recognition.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2016; pp. 770–778.
12. Perez L, Wang J: **The effectiveness of data augmentation in image classification using deep learning.** *arXiv preprint arXiv: 1712.04621.* 2017.
13. Rizve MN, Duarte K, Rawat YS, *et al.*: **In defense of pseudo-labeling: An uncertainty-aware pseudo-labeling framework for semi-supervised learning.** *arXiv preprint arXiv:2101.06329.* 2021.
14. Alam F, Ofli F, Imran M: **Processing social media images for humanitarian aid: A survey.** *arXiv preprint arXiv:2002.09394.* 2020.
15. Alam F, Ofli F, Imran M: **A survey on multimodal crisis informatics: A survey of datasets, methods, and platforms.** *arXiv preprint arXiv: 2103.01211.* 2021.
16. Liu Y, Ott M, Goyal N, *et al.*: **Roberta: A robustly optimized bert pretraining approach.** *arXiv preprint arXiv:1907.11692.* 2019.
17. Vaswani A, Shazeer N, Parmar N, *et al.*: **Attention is all you need.** *Advances in neural information processing systems.* 2017; pp. 5998–6008.
18. Chen Y, Marcus J: **The impact of online social networking on the self-concept of college students.** *J. Appl. Dev. Psychol.* 2012; **33**: 331–340.
19. Zignani M, Gaito S, Rossi GP: **A first look at the Fediverse.** *IEEE Internet Comput.* 2020; **24**: 23–31.
20. Boididou C, Papadopoulos S, Kompatsiaris Y, *et al.*: **Verifying information in social media.** *The SAGE handbook of social media.* Sage; 2018; pp. 323–338.
21. Tan M, Le Q: **Efficientnet: Rethinking model scaling for convolutional neural networks.** *International conference on machine learning.* PMLR; 2019; pp. 6105–6114.
22. **OpenAI: GPT-4 Technical Report.** *arXiv preprint arXiv:2303.08774.* 2023.
23. Hutson M: **Artificial intelligence faces reproducibility crisis.** *Science.* 2018; **359**: 725–726. PubMed Abstract | Publisher Full Text
24. Gundersen OE, Kjensmo S: **State of the art: Reproducibility in artificial intelligence.** *Proceedings of the AAAI conference on artificial intelligence.* 2018; Vol. **32**(1).
25. Stodden V, Leisch F, Peng RD: *Implementing reproducible research.* CRC Press; 2014.
26. Davenport SW, Bergman SM, Bergman JZ, *et al.*: **Twitter versus Facebook: Exploring the role of narcissism in the motives and usage of different social media platforms.** *Comput. Hum. Behav.* 2014; **32**: 154–162.
27. Howard J, Ruder S: **Universal language model fine-tuning for text classification.** *arXiv preprint arXiv:1801.06146.* 2018.
28. Lee DH: **Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.** *Workshop on Challenges in Representation Learning.* ICML; 2013; Vol. 3: p. 896.
29. Wei J, Zou K: **Eda: Easy data augmentation techniques for boosting performance on text classification tasks.** *arXiv preprint arXiv:1901.11196.* 2019.
30. Conneau A, Lample G, Rinott R, *et al.*: **Xnli: Evaluating cross-lingual sentence representations.** *arXiv preprint arXiv:1809.05053.* 2018.

31. Shorten C, Khoshgoftaar TM: **A survey on image data augmentation for deep learning.** *J. Big Data.* 2019; **6**: 1–48.

32. Van Engelen JE, Hoos HH: **A survey on semi-supervised learning.** *Mach. Learn.* 2020; **109**: 373–440.

33. Xie Q, Dai Z, Hovy E, *et al.*: **Unsupervised data augmentation for consistency training.** *arXiv preprint arXiv:1904.12848.* 2020.

34. Yun S, Han D, Oh SJ, *et al.*: **Cutmix: Regularization strategy to train strong classifiers with localizable features.** *Proceedings of the IEEE/CVF international conference on computer vision.* 2019; pp. 6023–6032.

35. Zhang H, Cisse M, Dauphin YN, *et al.*: **mixup: Beyond empirical risk minimization.** *arXiv preprint arXiv:1710.09412.* 2017.

36. Ata A, Kozak K: **Lightweight NSM - Narcissism Detection Dataset and Evaluation Results.** *Zenodo.* 2026.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✗

Version 1

Reviewer Report 08 April 2026

https://doi.org/10.21956/openreseurope.24046.r71370

© **2026 Tran H.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✗ **Hao Yen Tran**
Ho Chi Minh City Open University, Ho Chi Minh City, Ho Chi Minh, Vietnam

The topic of psychology has long attracted significant academic attention, particularly in an era where individuals are increasingly exposed to mental health challenges. Therefore, conducting in-depth and rigorous research in this domain is both important and necessary. While this topic is undoubtedly timely, the current study still suffers from several critical limitations that need to be addressed.

- First, the problem framing is rather superficial. The relationship between narcissism and social media has already been widely documented. However, the authors merely provide general background information and list research questions without clearly identifying a research gap or articulating the specific relationship they intend to investigate.
- Second, it remains unclear what distinguishes Mastodon from other social media ecosystems and why it deserves particular scholarly attention. Based on the current introduction, the novelty and unique contribution of the study are not convincingly demonstrated.
- Third, the related work section is also underdeveloped. The authors tend to make assertive statements without sufficiently guiding the reader through how this topic has been studied previously and why it matters specifically in the context of Mastodon. As a result, the significance of the research is not effectively communicated.
- Fourth, the statement *"Each post is then annotated for the presence of nine narcissistic traits based on the DSM-5 criteria"* lacks clarity. The paper does not explicitly define what these nine traits are, which makes it difficult for readers to assess the validity of the annotation process.
- Fifth, the most serious concerns lie in the methodological and data-related aspects of the study. The authors claim that they aim to address reproducibility issues by using a re-collectable dataset and a transparent modeling pipeline. However, in practice, both the data collection and filtering processes appear overly simplistic. Given that the study aims to assess psychological traits, this raises substantial concerns.
- Sixth, mental health is a sensitive domain. A methodology that directly engages with psychological constructs must adhere to stricter standards. Otherwise, there is a real risk of loosely labeling individuals as "narcissistic." Such labeling, even indirectly, may have unintended psychological implications, as being associated with narcissism can be personally and socially stigmatizing.

- Seventh, although the study is not conducted directly on human subjects, its implications clearly target human behavior. Therefore, a higher level of methodological rigor and ethical consideration is required. In particular, the authors need to clearly justify:
  - How the evaluation is conducted
  - How conclusions are derived from only two keywords
  - What criteria are used to filter posts
  - And on what basis a piece of content is considered “narcissistic”.
- Eighth, data-related limitations. The dataset suffers from several critical issues:
  - Sampling bias (restricted to only two hashtags)
  - Annotation validity (reliance on an LLM instead of domain experts)
  - Limited dataset size (2,000 samples)
  - Data mismatch (text-only analysis for inherently visual behavior)
- Last one, critical logical concerns. A major issue lies in the assumption that hashtags can serve as meaningful proxies for psychological traits. Using a hashtag such as #selfie does not imply that a user is narcissistic. It may simply reflect:
  - a trending behavior
  - a humorous or ironic expression
  - or casual, context-dependent usage

Is the work clearly and accurately presented and does it engage with the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research interests lie at the intersection of behavioral psychology and digital environments, with a particular focus on how underlying psychological constructs are manifested in social media behavior and consumer decision-making. I am especially interested in studies that examine the deeper psychological structures driving human behavior and how these mechanisms influence real-world outcomes.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 26 March 2026

https://doi.org/10.21956/openreseurope.24046.r71080

© 2026 Jornkokgoud K. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Khanitin Jornkokgoud
Mahasarakham University, Talat, Maha Sarakham, Thailand

Overall, the study presents an interesting exploratory computational approach to detecting narcissistic expressions on social media. Strengths include the use of DSM-5 criteria as a theoretical framework, the focus on an emerging platform (Mastodon), a transparent and reproducible lightweight machine-learning pipeline, and attempts to align linguistic features with clinical concepts such as grandiosity and need for admiration. However, the psychological validity of the narcissism construct remains limited, with potential confusion between narcissistic traits and personality disorder. The study would benefit from stronger theoretical justification and clearer validation of the construct being measured.

Major points
1. The use of DSM-5 criteria to classify individual social media posts raises concerns regarding construct validity. Narcissistic Personality Disorder criteria were designed for clinical assessment of stable personality patterns, not for evaluating isolated textual expressions. The authors should clarify that their model detects narcissistic expressions rather than narcissistic personality traits.

2. The decision to label posts as narcissistic based on ≥2 DSM-5 criteria are not theoretically justified. This threshold does not correspond to either dimensional trait models or clinical diagnosis and may lead to misclassification. This is an overinterpretation of hashtag differences. The authors claim that "#selfie posts were 7.5 times more likely to be narcissistic". It is methodologically weak because sampling restricted to only two hashtags, no control for user-level clustering, no personality validation, no demographic controls. This is behavioral proxy, not narcissism.

3.The study treats narcissism as a single construct. However, contemporary literature distinguishes between grandiose and vulnerable narcissism, or communal narcissism or genetic narcissism, which exhibit different linguistic patterns. The authors should discuss this limitation.

4. The dataset was annotated using GPT-4.1-mini without human validation. The absence of inter-rater reliability or clinician review raises concerns about label validity.

5. Personality traits are stable individual characteristics, yet the analysis is performed at the post

level. This conflicts with personality theory, including that narcissism is a stable trait and posts are situational behaviors. So, aggregating posts at the user level would better align with personality theory.

Minor issues
- Consider referencing dimensional models (e.g., narcissism spectrum)
- Clarify distinction between behavior and personality
- Discuss cultural differences in self-presentation

Is the work clearly and accurately presented and does it engage with the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

Are all the source data and materials underlying the results available?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Personality

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.