HippoFlow: Personalized Linguistic Embeddings for Predicting SMS Authorship and Personality Traits

Michael Angelo Robert Harrigan, Rhys Jones, Carissa Brinn Bostian
Virginia Tech

Introduction

Mental health crises affect over 3 million Americans seeking emergency psychiatric care annually; yet, no real-time monitoring systems exist. While previous NLP approaches to mental health analysis have relied on social media and therapy transcripts, SMS messaging – a more personal and frequent data source – remains understudied, with research limited to basic word frequency models.

We propose two transformer-based approaches for creating personalized linguistic embeddings from SMS history: (1) a conversation-agnostic bag of words approach using weighted word2vec embeddings, and (2) context-aware message-level transformer embeddings.

To validate these methods' potential for therapeutic applications – such as crisis prediction, optimal patient-therapist matchings, and personality assessment – we evaluate embedding user-uniqueness by comparing within-subject versus between-subject cosine similarities, and demonstrate effective message authorship prediction.

Methodology

Four participants aged 18-22 provided at least 27,000 SMS messages each, spanning approximately five years. This age range was selected due to individuals in this age range being at heightened risk of mental illness and their high messaging frequency.

Embedding Stability – BoW Implementation

- Embeddings for both 6-month bins and globally were created for each individual.
- word2vec word-embeddings were mean pooled over the appropriate temporal bin to create a single linguistic embedding for this timeframe.
- Cosine similarity was computed:
  - **Within-subject**: Each bin compared to the individual's full-history embedding.
  - **Between-subject**: Individuals' global embeddings were compared..
- Empirical bounds were used to test uniqueness of these embeddings.

Authorship Classification – Transformer Implementation

Model:

- Full fine-tuning of a lightweight transformer model to predict message authorship (n=4, chance=25%).
- Input: Individual messages.
- Output: Softmax classification over author IDs.

Experimental Design:

- Messages were labeled with true author ID.
- Per-author training size was scaled from 5 to 27,480 messages.
- Accuracy was computed on a held-out stable test set.

Results

Embedding Stability – BoW Implementation

We assessed the stability of the BoW embeddings by measuring the cosine similarity between message bins from the same user across time (within group)) and between different subjects (between group). Non-overlapping ranges in cosine similarities point evidence user-uniqueness in embeddings.

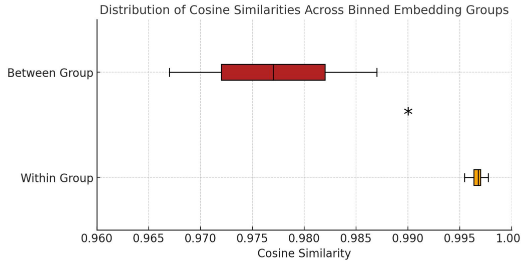

Figure 1: Within-subject vs. Between-subject Embedding Similarity. The non-existence of overlap between distributions demonstrates that our BOW embeddings capture user-specific linguistic patterns, which indicates that these embeddings effectively differentiate between users.

Authorship Classification – Transformer Implementation

We assessed how authorship classification performance scaled with training dataset size, from as few as 5 messages per user to 10,000 messages. Note chance level or 25% among n=4 participants.

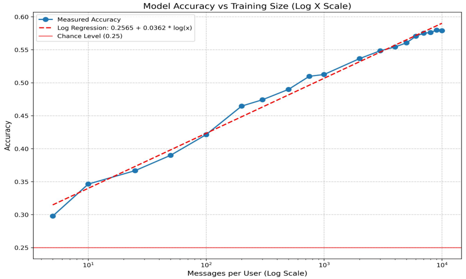

Figure 2: Scaling of Authorship Classification Accuracy with Training Size. Classification accuracy improves logarithmically with more messages per user, rising from ~30 % to ~59 % as data increases from 5 to 10 000 messages.

Discussion

This study demonstrates the ability of SMS messages to measure linguistic behavior over time in a stable manner and uniquely identifying individuals aged 18-22. The two complementary methods confirmed this:

1. *BoW-based Word2Vec* embeddings consistently differentiated between users with high within-subject similarity and lower between-subject similarity, despise the model's simplicity.
2. *Transformer-based classification* revealed that authorship classification of SMS messages via transformer full fine-tuning was significantly effective on the order of 50 labeled messages per user. Moreover, classification accuracy scaled logarithmically through 10,000 messages per user.

The non-overlapping distributions in cosine similarity underscore that users maintain persistent, individualized linguistic patterns over years of communication via SMS. Additionally, the logarithmic increase in classification accuracy with training size demonstrates a point of optimization that can be reached with increased data collection.

In all, we demonstrate that modern transformer-based personalized linguistic embeddings are unique to an individual. This finding emphasizes the feasibility of future architectures to predict more insightful personality insights.

Future Work

Upcoming efforts will apply SMS embeddings to prediction tasks related to mental health. This includes both classification (e.g., suicidality, mood states), and regression (e.g., Big Five traits, anxiety scores) models built atop the SMS embedding architecture. Moreover, we plan to employ stacked transformer architectures to whole conversations to embed context between messages.

Finally, while stability over the entire 5-year timeframe indicates predictive power of the embeddings, it raises questions over the ability to detect acute changes in personality, such as suicidal ideation. This homogeneity over time may be due to the large bin size of 6-mo. Binning at higher frequencies (e.g., 1-day as opposed to 6-mo intervals) may elucidate acute temporal changes in linguistic behavior.

References

1. Inkster, B., Stillwell, D., Kosinski, M., & Jones, P. B. (2022). Natural language processing applied to mental illness detection: A narrative review. npj Digital Medicine, 5, 1-13.
2. Tauscher, J. S., Lybarger, K., Ding, X., Chander, A., Hudenko, W. J., Cohen, T., & Ben-Zeev, D. (2022). Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. Psychiatric Services, 74(4), 409-416.
3. Cahn, D. (2021). DeepHelp: Deep learning for Shout crisis text conversations. arXiv preprint arXiv:2110.13244.
4. Ramezani, M., Feizi-Derakhshi, M.-R., & Balafar, M.-A. (2022). Knowledge graph-enabled text-based automatic personality prediction. arXiv preprint arXiv:2203.09103.
5. Peters, H., Cerf, M., & Matz, S. C. (2024). Large language models can infer personality from free-form user interactions. arXiv preprint arXiv:2405.13052.
6. Sim, K. Y. H., Fortuno, K. T., & Choo, K. T. W. (2024). Towards understanding emotions for engaged mental health conversations. arXiv preprint arXiv:2406.11135.