Supplemental Materials for Different Affordances on Facebook and SMS Text Messaging Do Not Impede Generalization of Language-Based Predictive Models

Tingting Liu[*,1,2] **Salvatore Giorgi**[*,1,2] **Xiangyu Tao,**[3] **Sharath Chandra Guntuku,**[2] **Douglas Bellew,**[1] **Brenda Curtis,**[†,1,‡] **Lyle Ungar**[†,2]

[1] National Institute on Drug Abuse
[2] University of Pennsylvania
[3] Fordham University
{tingting.liu, sal.giorgi, doug.bellew, brenda.curtis}@nih.gov, xtao16@fordham.edu, {sharathg, ungar}@cis.upenn.edu

Study Recruitment

Participants were recruited online via the Qualtrics Panel as part of a larger national survey on mental health, substance use, and COVID-19. To qualify, consenting participants must be 18 years or older, U.S. residents, and Facebook users. To ensure they are active Facebook users, participants must have posted at least 500 words across their lifetime status updates and posted at least 5 posts within the past 180 days to be included in the study (Eichstaedt et al. 2021). 2,796 participants finished an initial survey, including questions about socio-demographics, and physical and mental health (e.g., depression, stress, and life satisfaction). This pool of participants has been used to study loneliness and alcohol use (Bragard et al. 2022) and COVID-related victimization (Tao et al. 2023). The study received approval from the Institutional Review Board at our institution.

After completing this survey, participants were invited to install the open-source mobile sensing application AWARE (Ferreira, Kostakos, and Dey 2015; Nishiyama et al. 2020) on their mobile phones. This application collects mobile sensor information (e.g., movement, app usage, and keystroke data). A total of 300 participants installed and ran the AWARE app for 30 days. Keystroke data is only available for Android users (N = 192), thus we excluded 108 iPhone users. We only consider the Google, Verizon, and Samsung messaging apps as our keystroke text messaging sources, hereafter referred to as SMS data. 69 users who wrote less than 500 words within the 30-day study period were further excluded. Finally, since the models used in this study are trained on monolingual English, 3 participants were removed due to mostly Spanish status updates. Thus, 120 participants entered our data analysis ($M(SD)_{age}$ = 36.46 (9.74), range: 18-65-year-old; 69% female; the highest level of education: 57% have four-year Bachelor's degree or higher; household income: 49% $> \$60,000$).

[*]These authors contributed equally.
[†]Share the senior authorship.
[‡]Corresponding author.

Text Based Measures

Age and Gender We applied an age and gender predictive lexica (Sap et al. 2014) built over a set of Facebook users who self-disclosed age and gender and shared their Facebook status updates. The final model predicted age with a Pearson *r* = 0.86 and and binary gender with an accuracy = 0.90.

Depression This model (Schwartz et al. 2017) was built on roughly 28,000 Facebook users who consented to share their Facebook data and answered the depression facet of neuroticism in the "Big 5" personality inventory, a 100-item personality questionnaire (the International Personality Item Pool proxy to the NEO-PI-R (Goldberg et al. 1999)). This model resulted in a prediction accuracy (Pearson *r*) of 0.386. Please see the original paper for full details. The RoBERTa-based model has a 10-fold cross-validation predictive accuracy of Pearson *r* = 0.36.

Life Satisfaction This model was built on roughly 2,700 Facebook users who consented to share their Facebook data and answered a life satisfaction questionnaire (see Cantril's Ladder below; Jaidka et al. 2020). The model was built using a set of 2,000 LDA topics and produced a prediction accuracy of Pearson *r* = 0.26. The RoBERTa-based models have a 10-fold cross-validation predictive accuracy of Pearson *r* = 0.29.

Stress Similar to the depression and life satisfaction models, the stress model (Guntuku et al. 2019) was built over a set of Facebook users who answered Cohen's Perceived Stress inventory (see description below). Again, 2,000 LDA topics were used as features in a 10-fold cross-validation setup. The final model's accuracy was Pearson *r* = 0.31. The RoBERTa-based models have a 10-fold cross-validation predictive accuracy of Pearson *r* = 0.33.

Survey Based Measures

Depression Frequency of depression symptoms in the past two weeks are accessed via the 9-item Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, and Williams 2001;

e.g., “Little interest or pleasure in doing things”) with response options ranged from 0 (Not at All) to 3 (Nearly Everyday).

Life Satisfaction Life satisfaction is measured via Cantril’s Ladder, which asks respondents to identify their current step on a ladder with steps numbered from zero at the bottom to 10 at the top, with the top representing the the best possible life, and the zero representing vice versa (Cantril 1965).

Stress Stress is measured by Cohen’s Perceived Stress Scale (PSS; Cohen, Kamarck, and Mermelstein 1983). A sample item is “In the last month, how often have you been upset because of something that happened unexpectedly?” Response options ranged from 0 (Never) to 4 (Very often).

Keystroke Data and Text Cleaning

The AWARE mobile sensing app logs each non-password keystroke on Android phones across all apps (e.g., text messages and search engine entries). These logs are stored one character at a time and include modifications such as deletions and auto-correct. For example, if a user searched “Talyor Swift” in a search engine, AWARE would log separate database entries for “T”, “Ta”, “Tal”, etc. If the same user misspelled “Talyor” while typing, AWARE would also log the misspelling and the delete key; for example “T”, “Ta”, “Tai”, “Ta” (i.e., a *backspace* occurred), “Tal”, etc. This presents a unique challenge when dealing with possibly sensitive information.

While the main goal of cleaning Personal Identifiable Information (PII) is to enable non-trusted sources to access the collected data by removing PII, a secondary goal is to replace the PII with a tag indicating what kind of data has been removed to allow deeper analysis. Basic cleaning of each string was done in several stages. The first was to remove PII data that was structurally identified by the device itself as either a password field or a phone number. The second stage was to use *spaCy’s* Name Entity Recognizer (NER) and to replace all flagged entities with their category label. The third stage was to check against a list of common data formats using regular expressions using a modified version of CommonRegex[1]. We noted that these category labels were ignored by our tokenizer and not used in the downstream analyses in the present study.

Cleaning keystroke data which changes 1 character at a time; however, contains an extra challenge over standard complete string cleaning. Detection of partial PII data that doesn’t yet match a known form (but will eventually) is required. We accomplished this by rolling future data back through the previous data in two stages. The first stage was that, each time when the completion of a new token at the end of a string was detected, we applied the replacement information, or lack thereof, back through the previous strings until the beginning of that token (there may be incomplete tokens which match NER that were not necessary to replace based on subsequent characters). This allowed us to clean data that might be removed via deletion before the entry is complete. The second stage was once the whole entry was complete, we rolled all of the changed data back through all of the incomplete string items for this entry. This involved overlaying data replacement item information for individual strings that were wholly contained by the completed entry information, or where the replacement data fields only overlap, merging the possible replacement item information together to create a compound tag. This process was executed automatically on the study data, with no human intervention, so as to minimize the risk of leaking sensitive information. Finally, we noted that while we collected full keystroke data, only the final text data which was sent via SMS was analyzed (i.e., no partial text messages are considered).

[1]https://github.com/madisonmay/CommonRegex

Word Cloud Visualization

Figure 1 shows the 1-to-3 grams associated with each platform, visualized as a word cloud.

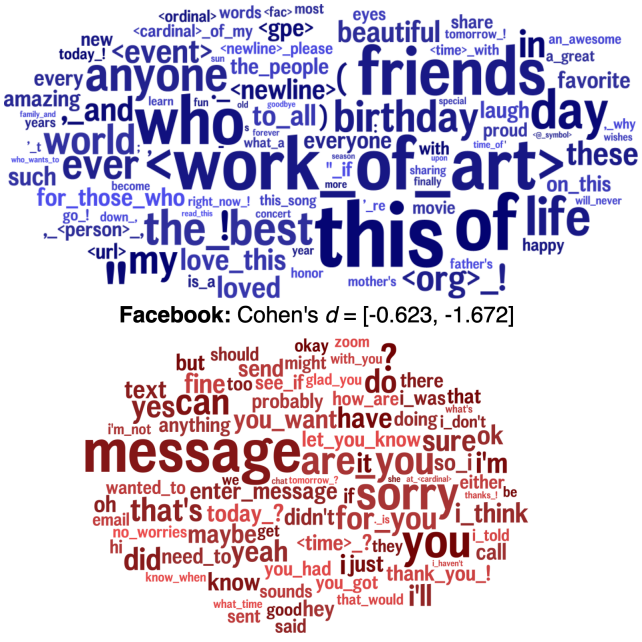

Figure 1: 1-to-3 grams most correlated with Facebook vs. SMS, statistically significant at $p < 0.05$ after Benjamini-Hochberg FDR correction. Cohen’s d = effect size measuring Facebook vs. SMS differences. N-grams size: larger more distinguishing; darkness: darker more frequent. Angle brackets: spaCy annotated named entities (e.g., <work of art>: titles of books, songs, etc).

References

Bragard, E.; Giorgi, S.; Juneau, P.; and Curtis, B. L. 2022. Loneliness and daily alcohol consumption during the COVID-19 pandemic. *Alcohol and Alcoholism (Oxford, Oxfordshire)*, 57(2): 198.

Cantril, H. 1965. *Pattern of human concerns*. New Brunswick, N.J.: Rutgers University Press.

Cohen, S.; Kamarck, T.; and Mermelstein, R. 1983. A global measure of perceived stress. *Journal of health and social behavior*, 385–396.

Eichstaedt, J. C.; Kern, M. L.; Yaden, D. B.; Schwartz, H.; Giorgi, S.; Park, G.; Hagan, C. A.; Tobolsky, V. A.; Smith, L. K.; Buffone, A.; et al. 2021. Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, 26(4): 398.

Ferreira, D.; Kostakos, V.; and Dey, A. K. 2015. AWARE: mobile context instrumentation framework. *Frontiers in ICT*, 2: 6.

Goldberg, L. R.; et al. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1): 7–28.

Guntuku, S. C.; Buffone, A.; Jaidka, K.; Eichstaedt, J. C.; and Ungar, L. H. 2019. Understanding and measuring psychological stress using social media. In *Proceedings of the international AAAI conference on web and social media*, volume 13, 214–225.

Jaidka, K.; Giorgi, S.; Schwartz, H. A.; Kern, M. L.; Ungar, L. H.; and Eichstaedt, J. C. 2020. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*.

Kroenke, K.; Spitzer, R. L.; and Williams, J. B. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9): 606–613.

Nishiyama, Y.; Ferreira, D.; Eigen, Y.; Sasaki, W.; Okoshi, T.; Nakazawa, J.; Dey, A. K.; and Sezaki, K. 2020. iOS crowd–sensing won't hurt a bit!: AWARE Framework and Sustainable Study Guideline for iOS Platform. In *International Conference on Human-Computer Interaction*, 223–243. Springer.

Sap, M.; Park, G.; Eichstaedt, J.; Kern, M.; Stillwell, D.; Kosinski, M.; Ungar, L.; and Schwartz, H. A. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1146–1151.

Schwartz, H. A.; Giorgi, S.; Sap, M.; Crutchley, P.; Ungar, L.; and Eichstaedt, J. 2017. DLATK: Differential language analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 55–60.

Tao, X.; Liu, T.; Fisher, C. B.; Giorgi, S.; and Curtis, B. 2023. COVID-related social determinants of substance use disorder among diverse US racial ethnic groups. *Social Science & Medicine*, 317: 115599.