Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

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

Abstract

Adaptive mobile device-based health interventions often use machine learning models trained on non-mobile device data, such as social media text, due to the difficulty and high expense of collecting large text message (SMS) data. Therefore, understanding the differences and generalization of models between these platforms is crucial for proper deployment. We examined the psycho-linguistic differences between Facebook and text messages, and their impact on out-of-domain model performance, using a sample of 120 users who shared both. We found that users use Facebook for sharing experiences (e.g., leisure) and SMS for task-oriented and conversational purposes (e.g., plan confirmations), reflecting the differences in the affordances. To examine the downstream effects of these differences, we used pre-trained Facebook-based language models to estimate age, gender, depression, life satisfaction, and stress on both Facebook and SMS. We found no significant differences in correlations between the estimates and self-reports across 6 of 8 models. These results suggest using pre-trained Facebook language models to achieve better accuracy with just-in-time interventions.

Introduction

Language reflects users' psychology and can be used to understand and predict mental health conditions (i.e., De Choudhury et al. 2013). While language from social media such as Facebook has been widely used (Eichstaedt et al. 2018; Jaidka, Guntuku, and Ungar 2018; Liu et al. 2022a), text messaging (Short Message Service or SMS) is emerging as a new platform for detecting mental health conditions and delivering interventions (e.g., depression: Liu et al. 2021, loneliness: Liu et al. 2022b). This also opens possibilities for Just-in-Time Adaptive Interventions (JITAIs) to deliver physical and mental health support based on an individual's changing state and environment (Nahum-Shani et al. 2018).

Most JITAIs are designed for smartphones, but current NLP models are primarily trained on social media, not SMS.

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

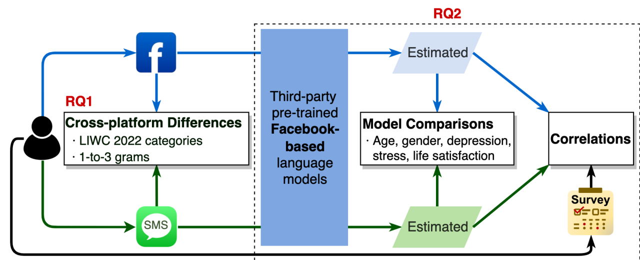

Figure 1: For the same users, we: (1) compared differences between Facebook and SMS language, and (2) evaluated the in- vs. out-of-domain efficacy of language models in predicting users' self-reported psychological traits.

Acquiring large-scale SMS for fine-tuning models is difficult and expensive. Transferring pre-trained Facebook models to SMS is thus common (Liu et al. 2021). But the impact of cross-platform language differences is unclear. Our evaluation aims to scientifically quantify the differences and ensure the successful transfer of current NLP models trained on social media sites into potential JITAIs using SMS.

Our research aims to explore: **RQ1** the distinctions in language between Facebook and SMS, and **RQ2** to evaluate the efficacy of language models derived from Facebook data in predicting psychological traits when applied to SMS. To achieve this, we utilize *the same cohort of users* who have provided their Facebook language, SMS, and psychometric self-reports (e.g., demographics, depression; Figure 1).

Contributions Our contributions are: (1) showing clear Facebook vs. SMS distinctions in language use for the same users; (2) evaluating the two platforms by training/validating within and across domains; and (3) laying the foundation for NLP model transfer to SMS with within-user comparisons.

Background

Language use varies across contexts. How can the language use of the same person differ in SMS and Facebook? Although, to our knowledge, no study has compared these two, some research has compared Facebook status updates with direct messages (DM), a private message on Facebook resembles SMS. Bazarova et al. (2013) observed that sharing positive emotions is associated with self-presentational concerns in Facebook status updates but not DM, noting the difference between communication on public and private channels. Bazarova and Choi (2014) also identified various goals and motivations for self-disclosures in Facebook status updates and DMs. Status updates relate to more social validation, self-expression, and relief, while DMs relate to relationship development and social maintenance.

If individuals use Facebook and SMS for different functions, it is still unclear how well models trained on Facebook posts will perform when applied to SMS data. Preliminary work showed that linguistic model predictions change across platforms. For example, Seabrook et al. (2018) examined the association between depression and emotional word expressions on Facebook and Twitter and found different patterns. On Facebook, the instability of negative emotion words only predicts depression, whereas on Twitter, the variability of negative emotion words reflects the severity of depression. However, as most cross-platform comparisons are made across user groups; individual differences (e.g., demographic) may cause these variations more than language choices. As cross-platform generalizations are expected to lead to model performance degradation, only a few studies have conducted same-user cross-platform comparisons (Jaidka, Guntuku, and Ungar 2018; Guntuku et al. 2019), no previous research has quantified these differences across Facebook and SMS within the same users. Our paper aim to draw more attention to these differences and provide possibilities for actionable improvements to conduct more precise predictions for future JITAIs plans.

Data

Participants Participants were recruited online via Qualtrics as part of a larger national survey (Tao et al. 2023). Each consenting participant (1) lived in the U.S., (2) was over 18 years old, (3) shared Facebook status updates, (4) installed the open-source mobile sensing application AWARE (Ferreira, Kostakos, and Dey 2015) on their Android phones, (5) wrote at least 500 words across platforms (Facebook and SMS apps), and (6) completed a survey which contains questions on age, gender[1], depression, life satisfaction, and stress. Our final sample included 120 participants (M_{age} = 36.46, 69% female).[2] Table 1 shows usage differences between Facebook and SMS.[3]

[1]We only analyze binary male/female gender,a limited and problematic sense of gender, due to limited data and the limitations of our gender estimation model. Three participants reporting a non-binary gender were excluded from the gender analysis.

[2]See Supplement for full details on participant recruitment, demographics, survey-based measures, and text-based estimates, at https://github.com/TTRUCurtis/Facebook-vs-SMS-language.

[3]Extensive cleaning was automatically applied (i.e., no human in the loop) to the keystroke data to remove any sensitive PII data. See Supplement for details. To fairly compare the Facebook data to the keystroke data, we applied the same cleaning pipeline to both.

	Words			Posts		
	Med.	Mean	SD	Med.	Mean	SD
FB	12,800	26,652	37,924	1,279	2,193	2,599
SMS	3,607	7,881	11,693	331	711	961

Table 1: Posts and word count statistics per platform (Med. = median and SD = standard deviation).

Survey-Based Measures For each participant, we collected self-reported age, gender, depression, stress, and life satisfaction via surveys used as gold-standard measures. We measured depression via the Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, and Williams 2001), life satisfaction via Cantril's Ladder (Cantril 1965), and stress via Cohen's Perceived Stress Scale (Cohen, Kamarck, and Mermelstein 1983).[2]

Text-Based Estimates We employed off-the-shelf text-based models to estimate age, gender (Sap et al. 2014), depression (Schwartz et al. 2017), stress (Guntuku et al. 2019), and life satisfaction (Jaidka et al. 2020). All models were developed in previous studies and trained on Facebook status updates to predict survey-based self-reports via lexical features (i.e., bag-of-words or bag-of-topics models).[2]

For this study, we also trained RoBERTa-based models (Liu et al. 2019) on the data sets used in the original papers. These models were trained for depression, life satisfaction, and stress only, as we cannot access the original data used to train the age and gender models. Since this paper is not aimed to build state-of-the-art classifiers, we used the same model pipeline across depression, life satisfaction, and stress: (1) we extracted user-level RoBERTa embeddings using the penultimate layer, (2) reduced the dimensions of the resulting 768 dimension embedding (using non-negative matrix factorization) to 128 dimensions (V Ganesan et al. 2021), and (3) applied a ℓ_2 regularized Ridge regression with $\alpha = 1$ (chosen via nested cross-validation). The RoBERTa-based models had similar accuracy to the lexical models.[2]

Methods

RQ1: Cross-platform Differences We first tokenized the Facebook status updates and SMS data, using a tokenizer designed for social media data (Schwartz et al. 2017). We considered both 1-to-3 grams and the Linguistic Inquiry and Word Count (LIWC) 2022 dictionary (Boyd et al. 2022). LIWC has been widely used in psychological sciences (e.g., Eichstaedt et al. 2018) and LIWC 2022 consists of 102 manually curated categories by psychologists. From both Facebook and SMS data, we extracted 1-to-3 grams and created a binary outcome variable for each participant to indicate which platform they were on. We then calculated effect size using Cohen's d values between platforms and conducted a logistic regression using n-grams to predict the binary platform indicator in order to calculate statistical significance (p values). Next, we extracted all LIWC 22 categories from each user's Facebook and SMS data. To calculate differences, we computed paired sample t-tests

Facebook		
Category	Top frequent words	*t*
Leisure	*fun, weekend, play*	14.65
Determiners	*the, a, my, this*	9.63
Quantities	*all, day, some, more*	7.91
Power	*own, order, power, president*	7.77
Emotion	*love, good, happy, :), fun*	7.35
SMS		
Category	Top frequent words	*t*
Auxiliary verbs	*is, have, be, was*	-20.26
Communication	*thank, say, thanks, said, tell*	-17.92
Discrepancy	*can, want, would*	-14.90
Assent	*yes, ok, yeah, okay*	-14.82
2nd person	*you, your, you're, u*	-13.27

Table 2: Paired t-tests results of LIWC 2022 categories, showing top categories which differ between Facebook and SMS. All results are statistically significant at $p < 0.001$ after Benjamini-Hochberg FDR correction.

for each LIWC category between Facebook and SMS. All significance thresholds were adjusted using a Benjamini-Hochberg False Discovery Rate (FDR) correction (Benjamini and Hochberg 1995).

RQ2: In vs. Out of Domain Estimates Here we performed three tasks to answer this from two different approaches: First, *Task 1* applied off-the-shelf models to both the Facebook and SMS data to evaluate in-and across-domain estimates and their generalization, and *Task 2* examined which linguistic features were driving the differences in estimates in *Task 1*. *Task 3* opted not to use off-the-shelf models in *Task 1* and *2*. Instead, it involved training and assessing predictive models within and across each domain.

Task 1: For each participant, we estimated age, gender, depression, life satisfaction, and stress from Facebook and SMS text using the text-based models described above. We then correlated the estimates with the gold-standard survey-based measures for both the lexical and embedding-based models. A statistical bootstrap test was used to assess differences in correlations between SMS-based estimates and Facebook-based estimates.

Task 2: To identify features driving lexical-based model estimates in both domains, we investigated feature importance i, which is defined as:

$$i(f) = w_f\big(freq_{FB}(f) - freq_{SMS}(f)\big). \qquad (1)$$

Here w_f is the weight of the feature f in the depression model, $freq_*(f)$ is the frequency of feature f in either the Facebook (FB) or SMS domain.

Task 3: Finally, instead of using off-the-shelf models, we trained and evaluated predictive models within and across each data set. To do this, we trained models to predict our five outcomes (age, gender, depression, life satisfaction, and stress) using both text sources from the same person as training and testing data sets: (1) train on FB / test on FB, (2) train on FB / test on SMS, (3) train on SMS / test on SMS, and (4) train on SMS / test on FB. We used a leave-one-out cross-

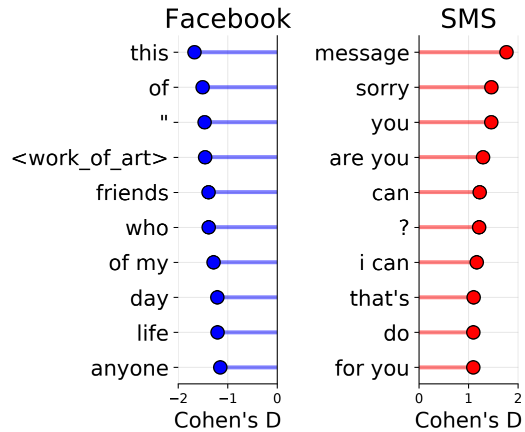

Figure 2: 1-to-3 grams most correlated with Facebook vs. SMS, statistically significant at $p < 0.05$ after Benjamini-Hochberg FDR correction. Cohen's d = effect size measuring Facebook vs. SMS differences. Angle brackets: spaCy annotated named entities (e.g., <work of art>).

validation setup when training and testing within the same text domain (1 and 3). When testing across text domains (2 and 4), we trained a model using one text source, applied the model to the other text source (producing estimates of our 5 outcomes), and correlated those estimates with self-reports.

Results

RQ1: Cross-platform Differences As seen in Table 2 and Figure 2, people preferred to discuss leisure activities, share pleasant feelings (LIWC positive emotion) and contents (e.g., books, songs), and express their motivations (LIWC power) on Facebook. People used more conversational language (LIWC communication), more second-person pronouns, and were more task-oriented (e.g., actions, LIWC verbs, plan confirmations) in SMS. See Supplement[2] for additional n-gram correlations that provided more context.

RQ2: In vs. Out-of-Domain Estimates In Table 3, we found that in-domain estimates from Facebook data predicted self-reports at rates similar to those in the original papers [2] from which the models were built (*Task 1*). When pre-

	Lexical Models		Embedding Models	
	Facebook	SMS	Facebook	SMS
Age	.68	.45*	-	-
Gender†	.91	.80*	-	-
Depression	.36	.29	.25	.08
Life Satis.	.21	.14	.31	.31
Stress	.21	.18	.21	.23

Table 3: Pearson correlations (or † accuracy) between language estimates and self-reports. * Significant difference in bootstrapping test between SMS and Facebook correlations.

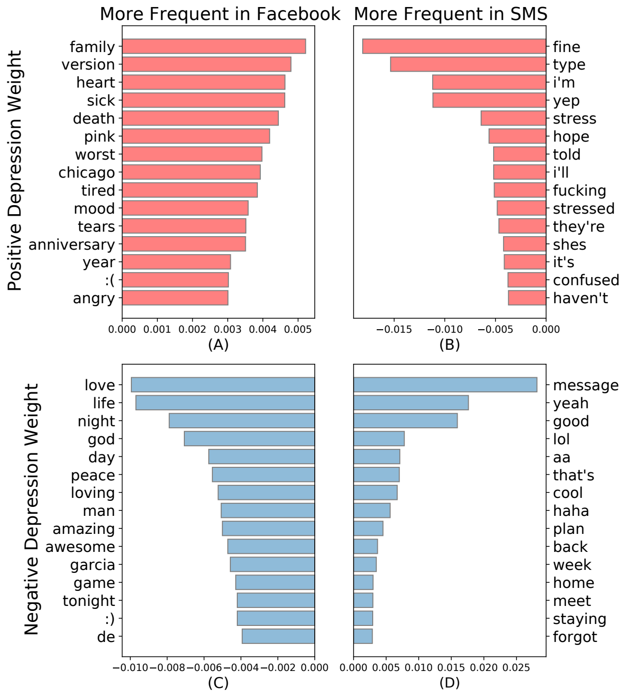

Figure 3: Feature importance results, as defined by the product of the depression model word weight and the difference in Facebook vs SMS word usage frequency. Top row (red bars; A and B) are positively weighted words in the depression model, while the bottom row (blue bars; C and D) are negatively weighted words. Left column (A and C) is more frequency words in Facebook (i.e., positive frequency difference), while the right column (B and D) contains words more frequent on SMS (i.e., negative frequency difference).

dicting self-reports from SMS-based estimates (i.e., out-of-domain), we observed a drop in prediction accuracy across all lexical models and 1 out of 3 embedding models. However, the differences between the Facebook correlation with self-report and the SMS correlation with self-report were not statistically different (using a bootstrapping test) except for those for age and gender (where SMS does not perform as well as Facebook). In Figure 3, we further investigated the drop in performance by examining feature importance (*Task 2*). Here we identified features reflecting language style, such as more use of contractions ("i'll", "i'm", "they're", "she's", "haven't"), driving the SMS depression estimates, and features about content, experience, and life events ("family", "sick", "anniversary") driving the Facebook depression estimates. In Table 4, we presented the results from training and evaluating models within and across domains. Facebook-trained models have higher in-domain accuracy, and SMS-trained models have higher out-of-domain accuracy. Again, using a bootstrapping significance test, we did not see significant differences between the correlation of Facebook and self-reports versus SMS and self-reports (in both in- and cross-domain tasks).

	In Domain		Cross Domain (train/test)	
	FB	SMS	FB/SMS	SMS/FB
Age	.61	.52	.40	.50
Gender†	.75	.74	.63	.73
Depression	.25	.09	.15	.32
Life Satis.	.19	.07	.25	.29
Stress	.24	.12	.32	.38

Table 4: Within and across platform evaluation. Pearson correlations (or † accuracy) between language estimates and self-reports. In Domain models are evaluated using leave-one-out cross validation.

Conclusion

Our study, based on data from *the same users*, shows: (1) individuals disclose different aspects of their lived experiences on Facebook and SMS, (2) two platforms generate similar mental health estimates, both within and across domains, whether using off-the-shelf models trained on Facebook data (Table 3) or models built specifically on the paper's dataset (Table 4). Consistent with past findings, Facebook usage reflects the need to belong and self-presentation (Nadkarni and Hofmann 2012), leading to more content sharing and opinion expression; whereas SMS is used for phatic communication to maintain social relationships and for informal discussions (Fibæk Bertel and Ling 2016), leading to more confirmations and conversational features. Our data, derived from the same users, indicates that cross-platform differences can be attributed to language rather than demographics. Despite the linguistic differences, our findings suggest that predictions from both platforms are similar.

Broader Impact

Our findings have important implications. Firstly, our research highlights the variations in psycho-linguistic features between Facebook and SMS, thus warranting further investigation of downstream applications. Secondly, future researchers can build predictive models on large-scale social media language and apply them to SMS, which may offer a new approach to address the cost-accuracy trade-off in the context of just-in-time interventions on mobile devices.

This study involves human subjects and was approved by the Institutional Review Board (IRB). The data used in this study raise ethical concerns such as handling sensitive personal information (PII) and thus, we have taken measures to securely store, clean, and analyze the data, further data sharing is not possible[3]. We use social media, SMS data, and machine learning methods to estimate sensitive attributes like depression. Such estimates can have both positive and negative implications, ranging from providing support to causing discrimination. We must use them with caution.

Acknowledgments

This study was supported by the Intramural Research Program of the NIH, National Institute on Drug Abuse (ZIA-DA000632). The authors report no conflict of interest.

References

Bazarova, N. N.; and Choi, Y. H. 2014. Self-Disclosure in Social Media: Extending the Functional Approach to Disclosure Motivations and Characteristics on Social Network Sites1. *Journal of Communication*, 64(4): 635–657.

Bazarova, N. N.; Taft, J. G.; Choi, Y. H.; and Cosley, D. 2013. Managing Impressions and Relationships on Facebook: Self-Presentational and Relational Concerns Revealed Through the Analysis of Language Style. *Journal of Language and Social Psychology*, 32(2): 121–141.

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300.

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*.

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.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Eichstaedt, J. C.; Smith, R. J.; Merchant, R. M.; Ungar, L. H.; Crutchley, P.; Preoţiuc-Pietro, D.; Asch, D. A.; and Schwartz, H. A. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44): 11203–11208.

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

Fibæk Bertel, T.; and Ling, R. 2016. "It's just not that exciting anymore": The changing centrality of SMS in the everyday lives of young Danes. *New Media & Society*, 18(7): 1293–1309.

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*.

Jaidka, K.; Guntuku, S.; and Ungar, L. 2018. Facebook versus Twitter: Differences in self-disclosure and trait prediction. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

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.

Liu, T.; Giorgi, S.; Yadeta, K.; Schwartz, H. A.; Ungar, L. H.; and Curtis, B. 2022a. Linguistic predictors from Facebook postings of substance use disorder treatment retention versus discontinuation. *The American Journal of Drug and Alcohol Abuse*, 48(5): 573–585.

Liu, T.; Meyerhoff, J.; Eichstaedt, J. C.; Karr, C. J.; Kaiser, S. M.; Kording, K. P.; Mohr, D. C.; and Ungar, L. H. 2021. The relationship between text message sentiment and self-reported depression. *Journal of affective disorders*.

Liu, T.; Ungar, L. H.; Curtis, B.; Sherman, G.; Yadeta, K.; Tay, L.; Eichstaedt, J. C.; and Guntuku, S. C. 2022b. Head versus heart: social media reveals differential language of loneliness from depression. *npj Mental Health Research*, 1(1): 16.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nadkarni, A.; and Hofmann, S. G. 2012. Why do people use Facebook? *Personality and individual differences*, 52(3): 243–249.

Nahum-Shani, I.; Smith, S. N.; Spring, B. J.; Collins, L. M.; Witkiewitz, K.; Tewari, A.; and Murphy, S. A. 2018. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6): 446–462.

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.

Seabrook, E. M.; Kern, M. L.; Fulcher, B. D.; and Rickard, N. S. 2018. Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. *Journal of medical Internet research*, 20(5): e9267.

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.

V Ganesan, A.; Matero, M.; Ravula, A. R.; Vu, H.; and Schwartz, H. A. 2021. Empirical Evaluation of Pre-trained Transformers for Human-Level NLP: The Role of Sample Size and Dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4515–4532. Online: Association for Computational Linguistics.