Towards LLM-Generated Affective Phrase Sets for Text Entry Evaluation

Andreas Komninos
University of Patras
Rio, Greece
akomninos@ceid.upatras.gr

Ioulia Simou
University of Patras
Rio, Greece
simo@ceid.upatras.gr

Angelos Fotopoulos
University of Patras
Rio, Greece
anfotopoulos@upatras.gr

Eleftheria Lito Michanetzi
University of Patras
Rio, Greece
st1067438@ceid.upatras.gr

Michalis Xenos
University of Patras
Rio, Greece
xenos@ceid.upatras.gr



Figure 1: Generated phrases across the Valence-Arousal space, from left to right V^+A^+; V^+A^-; V^-A^+; V^-A^-; V^0A^0 (neutral)

Abstract

We present a novel method to generate affective phrase sets using Large Language Models (LLMs), for use in text entry study transcription tasks. These phrases could be used to adjust participant emotions, and enable performance assessment under various affective conditions. We found that the affective phrases impacted participants' valence, but not arousal or text entry performance. Though partially succesful, our work provides a foundation for exploring the relationship between emotion and text entry performance in lab studies.

CCS Concepts

• **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; **Text input**.

Keywords

Text Entry, Smartphones, Transcription Tasks, LLMs, Phrase sets

ACM Reference Format:
Andreas Komninos, Ioulia Simou, Angelos Fotopoulos, Eleftheria Lito Michanetzi, and Michalis Xenos. 2024. Towards LLM-Generated Affective Phrase Sets for Text Entry Evaluation. In *International Conference on Mobile and Ubiquitous Multimedia (MUM '24), December 1–4, 2024, Stockholm, Sweden.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3701571.3703384

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
MUM '24, December 1–4, 2024, Stockholm, Sweden
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1283-8/24/12
https://doi.org/10.1145/3701571.3703384

1 Introduction

Mobile input methods (IMEs) support computer-mediated communications (CMCs), a core constituent of smartphone use. Novel IMEs are typically evaluated with lab-based transcription tasks (TTs), often using phrase sets derived from real CMCs (e.g. [13]). Actual CMCs can contain rich emotional signals [4], arising from the choice of words in the exchanged text, and cues such as capitalisation, emojis or even typos. Cognitive processing of these signals can impact the CMC recipient's affective state [2], possibly affecting their performance in the cognitive task of composing a response message [6, 7]. However, to our knowledge, the interplay of emotion and text entry performance has rarely been investigated to date [8, 9]. There are only a few examples in literature where the researchers attempted to control participants affective states during text entry, either with the intent of altering arousal (stress) through imposing unrealistic constraints to entry (e.g. timers), or by using methods beyond the transcription task [3, 5, 15]. Here, we present a method for generating *affective phrase sets* to use with TTs, aiming to enhance the TTs' ability to investigate text entry performance under various conditions in the valence-arousal (VA) space. Our aim is to reduce the problem of external validity of TTs [14] by integrating emotions into the task environment.

2 Generating Affective Phrases

As with real CMCs, we posited that the wording of phrase stimuli in TTs could alter participants' affective states. Using the VA classifier in [11] we explored one of the most commonly used phrase sets ([13]) and found that it does not equally represent all VA space regions (Fig. 2 left). Therefore, we propose a method for generating new affective phrase sets. We leverage the VA ratings of English words provided by two well-known lexicons [12, 16]. We normalise the word VA ratings in both lexicons between [0, 1]. Then, in the

2D VA space of each lexicon, we set up a grid of VA *anchor points* with a point spacing of 0.2 in both dimensions. Visiting each anchor point in succession, we obtain the k nearest neighbouring words (in this paper, $k = 20$). From these we select a random subset of 4 words, and instruct a state-of-the-art open-source LLM (in this paper, *Llama3.1-8B Q6K*) to generate a short phrase containing at least 2 of these words. The premise is that the generated phrase's VA rating *should* match the VA rating of the anchor point, since it contains words at or near that point in the VA space. We generated several thousands of samples on a commodity RTX-3070 desktop. The resulting set seems to gravitate towards the neutral space and resulting phrases do not always correspond to their originating anchor point quadrant (Fig. 2 right), but the generated set can be sampled, after eliminating phrases whose VA classification does not match the VA classification of the anchor point from which they were generated. We classify a dimension D in the VA space as negative (D^-) if $D \in [0, 0.4)$, neutral (D^0) if $D \in [0.4, 0.6]$ and positive (D^+) if $D \in (0.6, 1]$.

3 User Study and Results

We conducted a within-subject study with 13 university participants (age $\bar{x} = 22.84, \sigma = 3.72$, 5 female, English proficiency: 10 C2, 3 B2). Participants transcribed phrases from the centra (neutral) and four outermost VA regions (see Fig. 1 & Fig. 2 left, labelled regions). The study involved two sessions with a 30 minute break in between. In each session, participants began with a block of 5 neutral (V^0A^0) phrases, randomly picked from a sample of 45 phrases in that VA region, followed by two blocks of 15 phrases, randomly picked from a sample of 45 phrases from the other VA regions. The order of the two outermost region blocks was on a balanced Latin-square design. Prior to each phrase block, participants were shown a 1-minute music video clip from the DEAP set [10] to prime them for the VA quadrant of the phrases to be transcribed next, since congruent affective priming in lexical tasks is shown to improve perception of textual affective semantics [18]. This was important in order to negate effects from linguistic ambiguities or personal interpretations of texts, particularly since our participants were not native English speakers. After each block, participants completed a NASA-TLX survey and self-reported affective state while typing, using the Affective Slider scale [1]. We collected keyboard events, and physiological signals (EDA, and ECG). EDA sensors were placed on participants' non-dominant hand to prevent interference with typing. Participants typed seated using the index finger on a table-fixed Nokia 8.3, using GBoard with all predictive features disabled.

Self-reported *Valence* was statistically significantly different across conditions (Friedman $\chi^2 = 26.682, p < 0.001$). Pairwise post-hoc tests with Bonferroni correction confirm the differences are statistically significant between high and low valence conditions: T-test (V^+A^+, V^-A^+): t=5.969, p=0.0; Wilcoxon (V^+A^+, V^-A^-): Z=-2.982, p=0.003; T-test (V^+A^-, V^-A^+): t=5.312, p=0.0; Wilcoxon (V^+A^-, V^-A^-): Z=-2.831, p=0.005. The same applies to between the neutral condition and two other quadrants: T-test (V^+A^+, V^0A^0): t=4.137, p=0.001; T-test (V^-A^+, V^0A^0): t=-4.533, p=0.001. Differences in self-reported *Arousal* were not statistically significant (ANOVA $F = 1.192, p = 0.324$). Based on these results we did not expect to find statistically significant results on the EDA and ECG, since these biosignals are usually closer correlated to arousal. Indeed for both sensor types, there are no statistically significant differences in the mean value of the signals across conditions (EDA Friedman $\chi^2 = 8.062, p = 0.089$; ECG ANOVA $F = 0.057, p = 0.994$). Examining metrics from keyboard data (Fig. 3), we found statistically significant differences only in the use of backspaces (Friedman $\chi^2 = 10.548, p = 0.032$). Post-hoc pairwise tests with Bonferroni correction showed this difference is significant between high valence and the neutral condition only: Wilcoxon (V^+A^+, V^0A^0): Z=-2.831, p=0.005; Wilcoxon (V^+A^-, V^0A^0): Z=-2.867, p=0.004. Accordingly, participants did not report any statistically significant differences between the NASA-TLX factors, except for Effort (Friedman $\chi^2 26.335, p = 0.0$), between the low valence conditions and the neutral condition: Wilcoxon (V^-A^+, V^0A^0): Z=-3.113, p=0.002; T-test (V^-A^-, V^0A^0): t=5.375, p=0.0.

4 Conclusion

Our findings suggest that our *affective phrases* can impact participants' valence but not arousal. Text entry performance remains unaffected. However, high-valence phrases might improve participant mood and reduce experiment tedium, potentially aiding recruitment and attendance. Modifying arousal may require alternative approaches like time limits or gamification. We could also consider presenting phrases with other stimuli, e.g. photos, or sounds [17]. Our findings are limited by participant numbers and biosignals may provide further confirmation if analyzed more complexly. Further, we use the 8b version of Llama-3.1, therefore phrase quality might not be optimal. Our source code and data are openly available at https://github.com/komis1/llm-emotional-phrase-set, and readers may experience the text entry experiment online at https://tinyurl.com/y79tm44d.

References

[1] Alberto Betella and Paul F. M. J. Verschure. 2016. The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLOS ONE* 11, 2 (Feb. 2016), e0148037. https://doi.org/10.1371/journal.pone.0148037

[2] Hayley Blunden and Andrew Brodsky. 2021. Beyond the Emoticon: Are There Unintentional Cues of Emotion in Email? *Personality and Social Psychology Bulletin* 47, 4 (April 2021), 565–579. https://doi.org/10.1177/0146167220936054

[3] Matteo Ciman, Katarzyna Wac, and Ombretta Gaggi. 2015. iSensestress: Assessing Stress through Human-Smartphone Interaction Analysis. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. 84–91. https://doi.org/10.4108/icst.pervasivehealth.2015.259280

[4] Daantje Derks, Agneta H. Fischer, and Arjan E. R. Bos. 2008. The Role of Emotion in Computer-Mediated Communication: A Review. *Computers in Human Behavior* 24, 3 (May 2008), 766–785. https://doi.org/10.1016/j.chb.2007.04.004

[5] Marc Exposito, Javier Hernandez, and Rosalind W. Picard. 2018. Affective Keys: Towards Unobtrusive Stress Sensing of Smartphone Users. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, 139–145. https://doi.org/10.1145/3236112.3236132

[6] Surjya Ghosh, Shivam Goenka, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Representation Learning for Emotion Recognition from Smartphone Keyboard Interactions. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 704–710. https://doi.org/10.1109/ACII.2019.8925518

[7] Surjya Ghosh, Kaustubh Hiware, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Does Emotion Influence the Use of Auto-Suggest during Smartphone Typing?. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 144–149. https://doi.org/10.1145/3301275.3302329

[8] Satchit Hari, Ajay, Sayan Sarcar, Sougata Sen, and Surjya Ghosh. 2022. AffectPro: Towards Constructing Affective Profile Combining Smartphone Typing Interaction and Emotion Self-reporting Pattern. In *Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 217–223. https://doi.org/10.1145/3536221.3556603

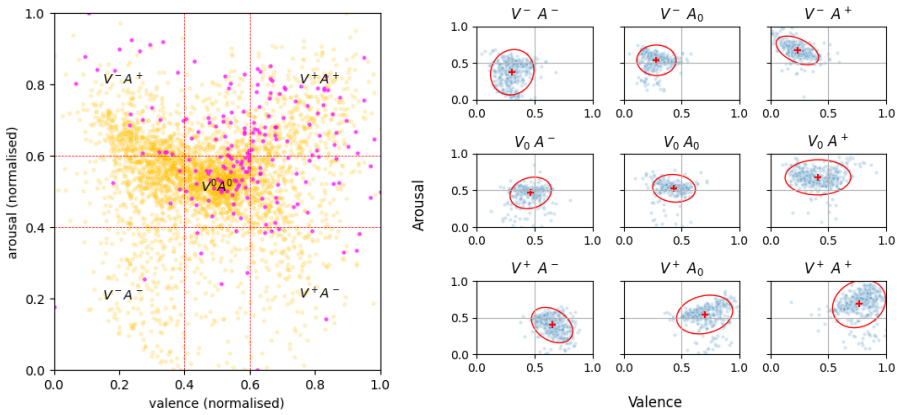


Figure 2: Left: Enron [13] (magenta) vs LLM-generated phrase VA loadings; Right: Dispersion of phrase VA loadings per generating anchor point region.

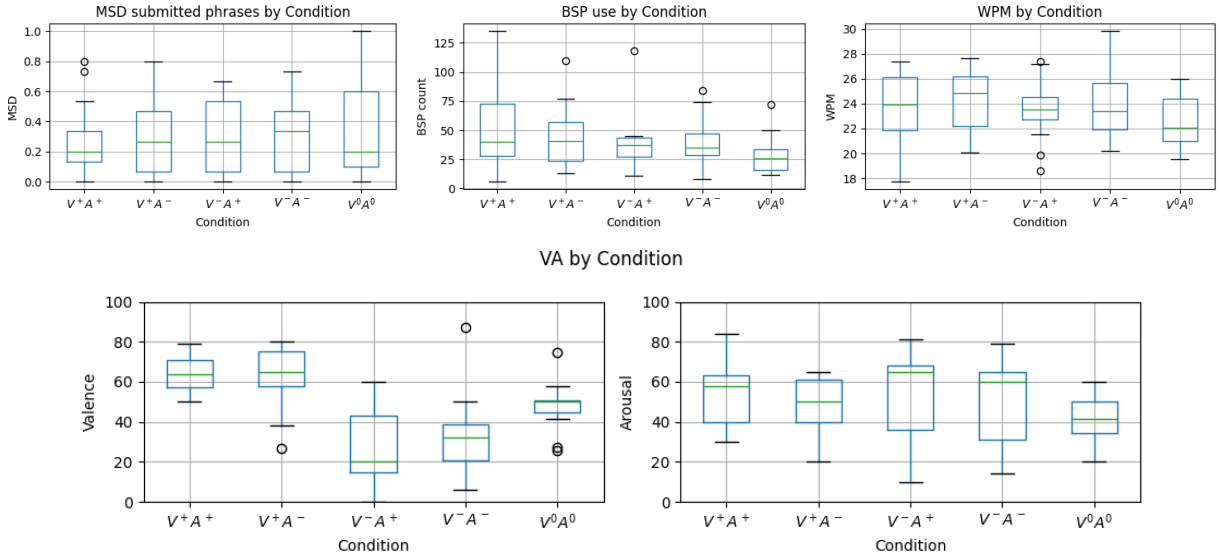


Figure 3: Results from keyboard and self-reported data.

[9] Susmi Jacob, P. Vinod, Arjun Subramanian, and Varun G. Menon. 2023. Affect Sensing from Smartphones through Touch and Motion Contexts. *Multimedia Systems* 29, 5 (Oct. 2023), 2495–2509. https://doi.org/10.1007/s00530-023-01142-6

[10] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (Jan. 2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[11] Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying Valence and Arousal in Text with Multilingual Pre-trained Transformers. In *Advances in Information Retrieval*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer Nature Switzerland, Cham, 84–100. https://doi.org/10.1007/978-3-031-28244-7_6

[12] Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 174–184. https://doi.org/10.18653/v1/P18-1017

[13] Keith Vertanen and Per Ola Kristensson. 2011. A Versatile Dataset for Text Entry Evaluations Based on Genuine Mobile Emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. Association for Computing Machinery, Stockholm, Sweden, 295–298. https://doi.org/10.1145/2037373.2037418

[14] Keith Vertanen and Per Ola Kristensson. 2014. Complementing Text Entry Evaluations with a Composition Task. *ACM Trans. Comput.-Hum. Interact.* 21, 2 (Feb. 2014), 8:1–8:33. https://doi.org/10.1145/2555691

[15] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, and Markus Gross. 2020. Affective State Prediction Based on Semi-Supervised Learning from Smartphone Touch Data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376504

[16] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior Research Methods* 45, 4 (Dec. 2013), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

[17] Liying Yang and Sheng-Feng Qin. 2021. A Review of Emotion Recognition Methods From Keystroke, Mouse, and Touchscreen Dynamics. *IEEE Access* 9 (2021), 162197–162213. https://doi.org/10.1109/ACCESS.2021.3132233

[18] Qin Zhang, Lingyue Kong, and Yang Jiang. 2012. The Interaction of Arousal and Valence in Affective Priming: Behavioral and Electrophysiological Evidence. *Brain Research* 1474 (Sept. 2012), 60–72. https://doi.org/10.1016/j.brainres.2012.07.023