Emotion Detection from Smartphone Keyboard Interactions: Role of Temporal vs Spectral Features

Salma Mandi
Indian Institute of Technology Kharagpur
INDIA
salmamandi@kgpian.iitkgp.ac.in

Surjya Ghosh
BITS Pilani Goa
INDIA
surjyag@goa.bits-pilani.ac.in

Pradipta De
Microsoft Corporation
USA
prade@microsoft.com

Bivas Mitra
Indian Institute of Technology Kharagpur
INDIA
bivas@cse.iitkgp.ac.in

ABSTRACT

Keystroke or typing dynamics represent two key facets - *rhythm* corresponds to spectral-domain characteristics and *timing* corresponds to time-domain behavior , which are created when a person types. The presence of inherent time-domain and frequency-domain characteristics in smartphone keyboard interactions motivate us to perform a comparative analysis of time-domain and frequency-domain features for emotion detection. We design, and develop an Android-based data collection application, which collects keyboard interaction logs and emotion self-reports (*happy, sad, stressed, relaxed*) from 18 subjects in a 3-week in-the-wild study. For the time-domain analysis, we extract a set of time-domain features and construct Random Forest-based personalized model; whereas for the spectral-domain analysis, first transform the interaction details into frequency-domain using DFT (Discrete Fourier Transform) and then extract a set of spectral-domain features to construct a personalized model for emotion detection. The empirical analysis from the study reveals that the time-domain models return superior classification performance (average AUCROC 72%) than the frequency-domain models (average 67%). It also signifies the importance of several time-domain and frequency-domain features as a strong discriminator of emotion states.

KEYWORDS

Emotion detection, Smartphone keyboard, Frequency Analysis, Fourier Transformation

ACM Reference Format:
Salma Mandi, Surjya Ghosh, Pradipta De, and Bivas Mitra. 2022. Emotion Detection from Smartphone Keyboard Interactions: Role of Temporal vs Spectral Features. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event, .* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3477314.3507159

SAC '22, April 25–29, 2022, Virtual Event,

ACM ISBN 978-1-4503-8713-2/22/04.
https://doi.org/10.1145/3477314.3507159

1 INTRODUCTION

Over the last decade, smartphone touchscreens became one of the most widely used input devices for user interaction [8, 9]. Recently, a significant volume of literature have explored the potential of the touch and keyboard interactions with the desktop keyboard or smartphone surfaces to infer user emotion, stress, and related affective states [1, 5, 7]. Those attempts mostly rely on the development of supervised machine learning (or deep learning) models, where suitable features are judiciously chosen to capture the signature of the different affect states. Hence, rigorous feature engineering is an integral component for the development of these models.

In this paper, we concentrate on the problem of emotion detection from the smartphone typing activities. However, exploration of the data collected from the typing modality opens up two avenues for the feature engineering (a) time domain features: which are simple to extract and have easy physical interpretation, (b) spectral or frequency domain features: which are obtained by converting the time based data into the frequency domain using the transformations. We collect typing interaction details in terms of elapsed time between two consecutive key presses (known as Inter-tap duration or ITD) from every *typing session*, defined as the time spent by the user at-a-stretch on a single application. We also collect the user's emotion self-report (from a set of four discrete labels - *happy, sad, stressed, relaxed*) corresponding to every session, by probing her at the end of the typing session. For the time-domain analysis, we extract several features from the set of ITDs and develop a machine learning model for emotion inference. On the contrary, for the frequency-domain analysis, we transform the session-wise ITDs to the spectral domain using Discrete Fourier Transform (DFT) and extract a set of features from the transformed data to train a machine learning model for emotion inference. Additionally, in order to examine the capacity of both temporal and spectral features together, we construct a hybrid model by aggregating the features from both the domains and compare its emotion classification performance (section 3). The experimental results demonstrate that using time-domain representations, user emotions can be detected more accurately with an average accuracy (AUCROC) of 72%, while in the frequency-domain the average accuracy (AUCROC) is 67%.

The major contribution of this paper is to highlight the potential of the typing based features for emotion classification. In this direction, we propose the time based and frequency based features

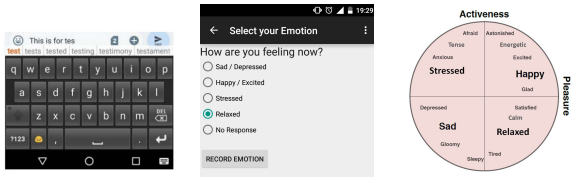
and developed simple ML based models (TDF and FDF) for emotion classification, which demonstrates the plausibility of the typing activities as emotion indicator.

2 DATA COLLECTION

In this section, first we describe the experiment apparatus, developed to collect the data. Next, we explain the data preprocessing steps and provide a brief summary of the collected data.

2.1 Experiment Apparatus

The experiment apparatus is an Android application, which consists of the following two key components - (a) app keyboard (Figure 1a), (b) emotion self-report collection UI (Figure 1b). We allow subjects to select one of the following four emotions (*happy, sad, stressed, relaxed*) as radio buttons. We choose these four emotions from four different quadrants of the Circumplex model (Figure 1c) of emotion [10], which allows user convenience in self reporting, since these four emotion labels are pretty discriminative & largely represent the frequently experienced emotion states [4]. The user also has the option to skip self-reporting by selecting the *No Response* option, which is set as default.



(a) App keyboard (b) Emotion self-report UI (c) Circumplex model [10]

2.2 Preprocessing & Data Summary

We recruited 18 participants (16 males, 2 females) aged between 24 to 33 years from our university. We installed the app on their smartphones and collected data for three weeks. We perform the data cleansing operations[5] on the collected data in the following three steps. (a) First, we remove all the sessions that are tagged with *No Response*, as they do not reveal any emotion (2.5% sessions). (b) Next, all the short sessions (with less than 80 typing interactions) are also eliminated, as lack of sufficient typing interactions may not be suitable for emotion prediction [5, 6] (22% sessions). (c) If the time interval between the end of a session and the collected emotion label is high, the staled label may not reflect accurately the emotion experienced during the session. Hence, we filter out all sessions for which the interval between the typing session and the emotion label collection is more than 1.5 hours (15% sessions).

Finally, after data preprocessing, we obtain a total of 504890 typing interactions spanning across 2100 typing sessions. On average we collect $\approx$ 90 sessions from every subject. We observe that all subjects have reported at least 3 emotions and all but 4 subjects (U10, U13, U14, U15) have reported all the four emotion labels. It is also observed that for most of the subjects *relaxed* emotion is most frequently reported, followed by *happy*, *stressed* and *sad*, which causes imbalance in the distribution of emotions.

3 EMOTION PREDICTION MODELS: TEMPORAL VS FREQUENCY

In this paper, we consider that the typing behavior of the subject carries the signature of her emotion states. We measure typing speed as Inter-Tap Duration (ITD), which is the elapsed time between two subsequent typing event. We define typing session as the time period, when subject stays onto a single application without changing the same. For example, when a subject uses WhatsApp uninterrupted without switching to other application from time t_1 to t_2, then we define elapsed time between t_1 and t_2 as a typing session. Each small bar within this session (see Figure. 2) depicts a typing event and we calculate the Inter-Tap Duration (ITD) as the interval between two consecutive typing events. We represent a session t of dimension n as a sequence of ITDs $S_n^t\{v_i|i \in \{1,\ldots,n\}\}$, where each element v_i in this vector refers the time interval between two consecutive typing events in a session.

3.1 Feature construction

3.1.1 Time domain feature. First, we compute the time domain features from the typing characteristics, obtained from the session t. We concentrate on the most straightforward features, which can be directly computed from the ITD sequence S_n^t of a session t. For instance, we compute simple statistic of S_n^t, such as mean, first quartile, second quartile, and third quartile from the ITDs in a session t. For each session of a subject, the mean session ITD is calculated as follows

$$MSI_{session} = \frac{\sum_{i=1}^{n} v_i}{n},$$

where n is length of session and v_i is ith ITD value

In order to compute the first, second, and third quartile of a session t, we first derive the sorted sequence of ITD values $S^t{}'_n = \{v_i|i \in \{1,\ldots,n\}\}$ of session t. We compute the second quartile of t as the median of the complete sorted sequence $S^t{}'_n$, whereas the first and third quartile are the median of lower 50% and upper 50% of $S^t{}'_n$ respectively. We extract a set of features from every session of a subject and use them to train the model.

3.1.2 Frequency domain feature. First, we apply Discrete Fourier Transform (DFT) on the ITDs present in a session t to obtain the equivalent frequency-domain representation. Each element in S_n^t, say v_i represents the Inter Tap Duration (ITD), which is a discrete element showing the time interval between two typing events. After transforming in the frequency domain, we represent this ITD sequence of session S_n^t (of dimension n) as a combination of n number of periodic signals with different amplitudes and frequencies. Hence, in the frequency domain, the obtained signals can be represented as a collection of complex numbers $\{x_{k_j} + iy_{k_j}, \forall j \in \{1,\ldots,n\}\}$. Here each real component x_{k_j} represents the amplitude of the respective signal with frequency k_j. Since we focus on the amplitude only, we discard the imaginary part and deal with only the real part of the coefficients [3]. Finally, we compute the resultant amplitude x_k of the session S_n^t for the signal with frequency k Hz as follows,

$$x_k = \sum_{j=0}^{n-1} v_j \times cos(2\pi k * j/n) \quad (1)$$

For a session S_n^t of dimension n, we repeat this procedure for all the n signals to generate the amplitude vector $A_n = \{x_{k_j}, \forall j \in \{1, \ldots, n\}\}$. We consider A_n^t as a frequency domain representation of session S_n^t, of dimension n. Note that depending on the session dimension, the cardinality of the amplitude vector may vary across various sessions.

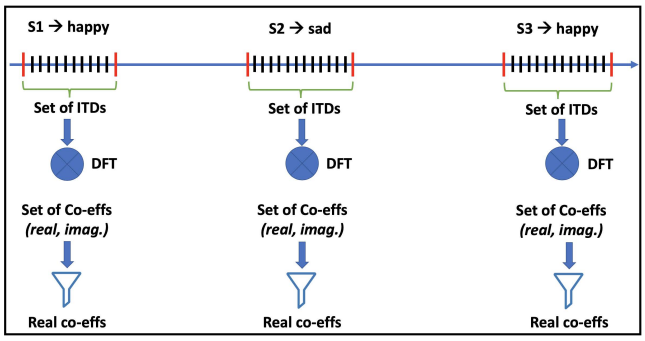


Figure 2: Schematic diagram of transforming set of ITD(s) to frequency domain. The set of ITD(s) obtained from different sessions (e.g. S1, S2, S3) are transformed using Discrete Fourier Transform (DFT) to obtain the frequency domain representation. From this set of coefficient-pairs, we filter out the imaginary ones and consider only the real ones for future processing.

3.2 Model implementation

We consider the top-3 peak amplitudes, extracted from the amplitude vector A_n^t of session t, as frequency-domain feature. Leveraging the time domain features and frequency domain features, we construct random forest-based personalized emotion prediction models, Time domain feature based model (TDF), and Frequency domain feature based model (FDF), respectively. There is no specific value chosen for the tree's maximum depth, so it is unlimited. For each emotion label (say, happy), we exclusively train a classifier model based on that label (say, happy) and rest of the labels (sad, relaxed, stressed). For each subject, each model is build separately to capture the personality trait.

Apart from individually exploring the capacity of the time domain and the frequency domain features, finally, we investigate the discriminating power of the time-domain and frequency-domain combined features. We develop a Hybrid model (HM) leveraging the features from both the time and frequency-domain.

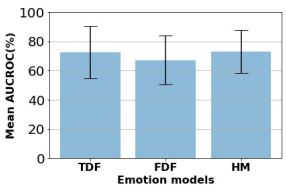
4 EVALUATION

In this section, we conduct a comparative study of the emotion prediction models, developed based on the temporal features (TDF), spectral features (FDF) and the combined features (HM) respectively.

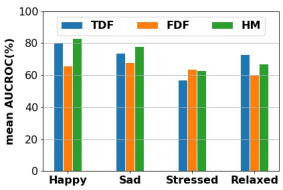
4.1 Experiment setup

We implement a personalized machine learning model for every user to classify the four emotion states. We implement stratified train-test split for every user to evaluate the performance of the developed models. We split the dataset in 80 : 20 ratio where 80% data is used to train the models and 20% is used for testing purpose. Then we apply Synthetic Minority Oversampling Technique (SMOTE)[2] only on the training data to inflate & balance the training dataset, whereas we test the models on the original data. The upsampling is done in such a way that the number of samples in all class are the same as the number of samples in the major class.

We compute unweighted AUCROC (Area under the Receiver Operating Characteristic curve) as the performance metrics, which compare the model predicted emotion, with the ground truth emotion labels. In order to compare the subject-wise accuracy, we implement the average of unweighted AUCROC (auc_{avg}) from four different emotion states.



(a) Model-wise mean AUCROC.



(b) Emotion-wise mean AUCROC.

Figure 3: Comparison of model-wise and emotion-wise performance. The TDF model performs better than the FDF model. Combining time-domain, and frequency-domain features performs similarly as the TDF model. Error bar indicates standard deviation.

4.2 Subject wise performance

In Figure 4 we compare the subject wise accuracy of the TDF, FDF and HM models. The TDF model provides an average auc_{avg} of 72% (std. dev 17%), while the FDF model obtains an average auc_{avg} of 67% (std. dev 16%). However, when both the temporal and spectral features are combined in the HM model, it performs similarly as the TDF model, as it achieves an average accuracy auc_{avg} of 73% (as shown in Figure 3a). In TDF, we observe that 50% of the subjects have an auc_{avg} of at least 70%, whereas in FDF and HM achieves an auc_{avg} of at least 60% and 70%, respectively. In case of HM, it is observed that around 77% of the subjects achieve an accuracy (auc_{avg}) of at least 60%, whereas it achieves 66% of the subjects in TDF. We achieve a 93%, 75%, and 93% accuracy (auc_{avg}) in TDF, FDF, and HM model, respectively for subject 7 (U7). However, it is observed that 100% accuracy(auc_{avg}) is obtained for subject 8 (U8), the accuracy is mean of accuracy for two emotion classes (*happy* and *relaxed*) as no sample from *stressed* and *sad* was present in the test set. Furthermore, it is impossible to apply SMOTE on the data of U8 due to the presence of very few data sample in some classes. We obtain poor classification performance for a few subjects (like 10, 11, 13, 14), primarily due to the skewed distribution of emotions in their dataset.

4.3 Emotion wise performance

We report emotion-wise accuracy for each model in Figure 3b. In all the models, the emotion state *happy* is classified with the highest AUCROC of 80%, 66%, and 83% in the TDF, FDF, and HM model,

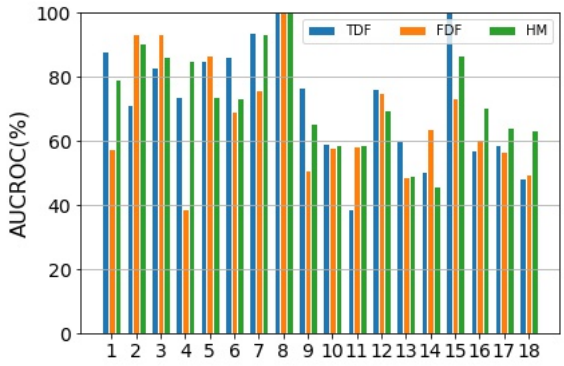


Figure 4: Subject-wise AUCROC for all models.

respectively. Although emotional state *sad* is the less frequent emotion among users, it results in the high AUCROC score due to its distinctive pattern than other emotions. On the other hand, emotion *stressed* is identified with the lowest accuracy of 57%, 64%, and 63% in the TDF, FDF, and HM model, respectively.

In FDF model, the *stressed* and *relaxed* states are identified with poor AUCROC scores indicating that frequency-domain features are not suitable to identify these two states. If we closely observe the data distribution of subjects, the highest number of samples is in emotion relaxed, followed by *happy*, *stressed*, and *sad*. But the number of the sample have no impact on the accuracy of the states. Instead, we observe that the emotional state *relaxed* is hard to distinguish as indicated by the accuracy.

4.4 Feature Importance: Temporal vs Spectral

We rely on the HM model to compute the *information gain (IG)* of the features by implementing *InfoGainAttributeEval* method from Weka [11]. In Table 1, we rank all the time-domain and frequency-domain features based on the average information gain. It is observed that second quartile (Q2) tops the list followed by the features first (Q1) and third (Q3) quartile. We notice all together top-3 amplitude have a moderate impact on models' performance. This suggests that easy-to-compute time-domain features are strong discriminator of emotion states than the frequency-domain features.

Features	Rank	Information Gain (IG)
Second quartile (Q2)	1	0.278
First quartile (Q1)	2	0.242
Third quartile (Q3)	3	0.189
Second Peak Amplitude (PA_2)	4	0.166
First Peak Amplitude (PA_1)	5	0.151
Third Peak Amplitude (PA_3)	6	0.148
MSI	7	0.106

Table 1: Ranking time-domain and frequency-domain features based on information gain. All top-3 features are from the time-domain, suggesting time-domain features are better discriminator of emotions than the frequency-domain features.

5 CONCLUSION

In this paper, we perform a comparative analysis between the time-domain and frequency-domain representation of smartphone typing interaction data for multi-state emotion detection. We design and develop an Android application (consisting of a keyboard and self-report collection facility), which is used to collect typing interaction details and emotion self-reports (*happy, sad, stressed, relaxed*) from 18 subjects in a 3-week in-the-wild study. We group the typing data in sessions, consisting of Inter-tap duration (ITDs) and labeled by the emotion self-reports. To perform the time-domain analysis, we extract a set of features from these ITDs and train a personalized Random Forest-based model for emotion classification. On the contrary, to perform the frequency-domain analysis, we transform the ITDs using DFT and extract a set of amplitude related features to train another personalized model for emotion detection. We obtain superior performance for the time-domain models and obtain an average AUCROC of 72%. For both time-domain and frequency-domain representations, several easy-to-compute features like (first, second, and third) quartiles of ITDs and (first, second, and third) peak amplitude of transformed ITDs are found to be strong discriminator of different emotions.

REFERENCES

[1] Bokai Cao, Lei Zheng, Chenwei Zhang, Philip S Yu, Andrea Piscitello, John Zulueta, Olu Ajilore, Kelly Ryan, and Alex D Leow. 2017. Deepmood: modeling mobile phone typing dynamics for mood detection. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 747–755.

[2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[3] M. Garrido, K. K. Parhi, and J. Grajal. 2009. A Pipelined FFT Architecture for Real-Valued Signals. *IEEE Transactions on Circuits and Systems I: Regular Papers* 56, 12 (2009), 2634–2643.

[4] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017. Evaluating effectiveness of smartphone typing as an indicator of user emotion. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 146–151.

[5] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2017. TapSense: Combining Self-report Patterns and Typing Characteristics for Smartphone Based Emotion Detection. In *Proceedings of the ACM MobileHCI*.

[6] S. Ghosh, N. Ganguly, B. Mitra, and P. De. 2019. Designing An Experience Sampling Method for Smartphone based Emotion Detection. *IEEE Transactions on Affective Computing* (2019), 1–1. https://doi.org/10.1109/TAFFC.2019.2905561

[7] Md Rakibul Islam, Md Kauser Ahmmed, and Minhaz F. Zibran. 2019. MarValous: Machine Learning Based Detection of Emotions in the Valence-Arousal Space in Software Engineering Text. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (Limassol, Cyprus) *(SAC '19)*. Association for Computing Machinery, New York, NY, USA, 1786–1793. https://doi.org/10.1145/3297280.3297455

[8] Huy Viet Le, Sven Mayer, Patrick Bader, and Niels Henze. 2018. Fingers' Range and Comfortable Area for One-Handed Smartphone Interaction Beyond the Touchscreen. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Sven Mayer, Huy Viet Le, Markus Funk, and Niels Henze. 2019. Finding the Sweet Spot: Analyzing Unrestricted Touchscreen Interaction In-the-Wild. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*. 171–179.

[10] James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[11] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.