"EMOTIONAL INSIGHTS AND PREDICTIONS USING NON-INTRUSIVE SMARTPHONE ACTIVITIES"

Research Paper

Barath Raj Kandur Raja, Samsung R&D Institute, Bangalore, India, barathraj.kr@samsung.com

Sumit Kumar, Samsung R&D Institute, Bangalore, India, sumit.kr@samsung.com

Mario Silic, Swiss School of Business and Management, Geneva, Switzerland, mario@ssbm.ch

Shwetank Choudhary, Samsung R&D Institute, Bangalore, India, sj.choudhary@samsung.com

Harichandana B S S, Samsung R&D Institute, Bangalore, India, hari.ss@samsung.com

Likhith Amarvaj, Samsung R&D Institute, Bangalore, India, likhith.amar@samsung.com

Latika Jindal, Samsung R&D Institute, Bangalore, India, latika.jdl@samsung.com

"Abstract"

Emotion identification is a complex research area that can enable unique multi-device experiences. Smartphones, the dominant mode of communication, can aid in emotion prediction. However, there is a lack of datasets with precise ground truth labels based on user smartphone behavior due to challenges associated with dataset annotation. Present annotation techniques rely either on self-reporting or recording on desktop applications, which is less natural. In this work, we address these issues by devising a user-centric approach to collect and annotate user data in a non-intrusive way on smartphones. We derive insights from the annotated data comprising behavior, emotion, and personality. The data consists of categorical features that do not include personally identifiable information, thus preserving user privacy. We validate the annotated data by an emotion prediction model using the Random Forest classifier, achieving an accuracy score of 67.73%. Further, we achieve an accuracy of 77.95 % on sentiment prediction (positive, negative, and neutral) using the Support Vector Machine (SVM) classifier.

Keywords: Emotion prediction, Data annotation, Privacy, Behavior Insights.

1 Introduction

Since the 17th century, the word "emotion" has been used in English as a translation of the French word "´emotion". The term began to be used considerably more frequently in 18th century English, often to refer to mental experiences. In the next century, it ultimately developed into a theoretical term due to the significant influence of two Scottish philosopher-physicians, Thomas Brown and Charles Bell (Dixon, 2012). Over the period, multiple researchers tried to develop models for understanding emotion. Broadly, two types of emotion models are considered: categorical and dimensional. Categorical emotion models divide emotions into distinct categories. Ekman's (1992) classic six-basic-emotion model comprising of emotions: happy, sad, anger, fear, surprise, and disgust is an example of the same. Robert Plutchik (2001) proposed another prominent categorical emotion model that described emotions as a combination of eight basic emotions. Compared to Categorical Models,

Dimensional Models express emotions using a multi-dimensional space. The Pleasure-Arousal-Dominance model (Russell et al., 1977) is one such dimensional model that argues that three dimensions of pleasure-displeasure, arousal-nonarousal, and dominance-submissiveness are sufficient to describe a large variety of emotional states.

Contrary to emotion, which is short-lived, personality traits are the cognitive, behavioral, emotional, and motivational characteristics of an individual that are more consistent across situations (McAdams, 2015). Each person varies along a spectrum on any given trait (e.g., a person can be open-minded, close-minded, or in between). Over the years, many theories have suggested the number of traits that can describe any individual's personality. There are different taxonomies discussed in the literature (Oliver et al., 1999). In this paper, we have focused on the most widely used big five human personality traits: extraversion, openness, conscientiousness, agreeableness, and neuroticism (McCrae et al., 2008).

Few attempts have been made to establish a relationship between personality and emotion, both equally important to understand the user. The Personality Emotion Model (PEM) workflow enables bi-directional personality-emotion mappings (Donovan et al., 2021). Sadeghian et al. (2021) shows that for emotional profiling or user behavior understanding, the personality can help in boosting the emotional model's accuracy. Our work attempts to build over this understanding and provide a more practical and non-intrusive way to understand this relation.

To study this relation, we opted to observe user behavior on the smartphone. People are spending a third of their waking time on smartphones as per app monitoring firm App Annie (data.ai, 2022), which is close to 4.8 hours. Thus, it is essential to understand the factors that can influence a user's emotions in a non-intrusive manner. Another important aspect while researching the relation is to maintain user privacy. Privacy of personally identifiable information (PII) and other user data is one of the significant concerns that permeate recent technological developments and associated regulations (Pelteret et al., 2016). Most of the previous works involving the task of emotion prediction from Smartphone usage utilize one or more PII, which may not be desirable to the user.

Considering all the above factors, the main contributions of the paper are manifold:

1. To the best of our knowledge, this is the first work to capture the relation between personality, emotion, and user smartphone behavior in a non-intrusive way, maintaining user privacy.
2. We propose a novel annotation technique using the smartphone for generating ground truth labels. We developed a user trial app based on the same and distributed it to over 100 participants for collecting tagged ground truth data.
3. We derive critical insights from the data collected from the user trial app about emotion, personality, and user behavior.
4. We validate the collected non-intrusive data by developing and implementing a system for automatic emotion detection. We extract different features from the collected tagged data, train the machine learning model and test its performance. Our system can detect emotion with an accuracy of 67.73% and sentiment with 77.95%, respectively.

2 Related Works

Data collection is the first and one of the most crucial steps in performing any analysis. Especially in the field of Machine learning (ML) and Artificial Intelligence (AI), the quality and depth of data will determine the level of AI applications that can be achieved. Data collection and analysis methods have been explored since the early 1600s when John Graunt (1977) conducted data analysis on gender-based death rates and attempted to predict life expectancy. Over the years, the research in this field evolved where questionnaire-based data collection was explored (Baker, 2003). Recently, due to the advent of smartphone usage, UI designs for more effective data collections have been extensively studied (Schobel, 2014).

Human emotion recognition is vital in improving the overall user experience. Emotion recognition has been pursued utilizing inputs from a variety of domains. Audio emotion recognition systems (Ooi et al., 2014) extract features like Mel Frequency Cepstral Coefficients (MFCC), Log-mel spectrograms, etc. from signals to determine the emotions. Another widely researched area is visual features-based emotion recognition. Vision-based systems like (Akhand et al., 2021) use facial expression features to detect the subject's emotion in the image. Studies like Quanzeng et al. (2016) focus on identifying the emotions evoked due to a particular image for the spectator. Other popular modalities from which human emotion is recognized is sensor-based techniques (Kanjo et al., 2016) and text (Majumder et al., 2019). The abundant availability of data also follows the wide research in these fields. Some of the popularly used datasets include MELD (Poria et al., 2018), IEMOCAP (Busso et al., 2008), RAVDESS (Livingstone et al., 2018) and many more.

With the advent of smartphone usage, one of the most significant indicators of a user's emotional state is the user-activity data. Although extensive research in emotion recognition utilizes various modalities as listed above, research in this field of predicting the mood/emotion of users using mobile-activity data, which is mostly categorical, is sparse. One important blocking factor is the lack of data. Previous work in this area includes MoodScope (LiKamWa et al., 2013). Here they create a data collection application and collect information like SMS, email, phone call, application usage, web browsing, and location from 32 participants four times a day to build statistical usage models to estimate mood. They use least-squares multiple linear regression to perform the modelling, along with Sequential Forward Selection (SFS). Bogomolov et al. (2014) and Hung et al. (2016) focus on detecting negative emotions like stress, depression from user activity data collected using a custom application. Since labelled data of this kind is sparse, iSelf (Sun et al., 2017) attempts to tackle the problem of cold- start conditions to predict emotion labels.

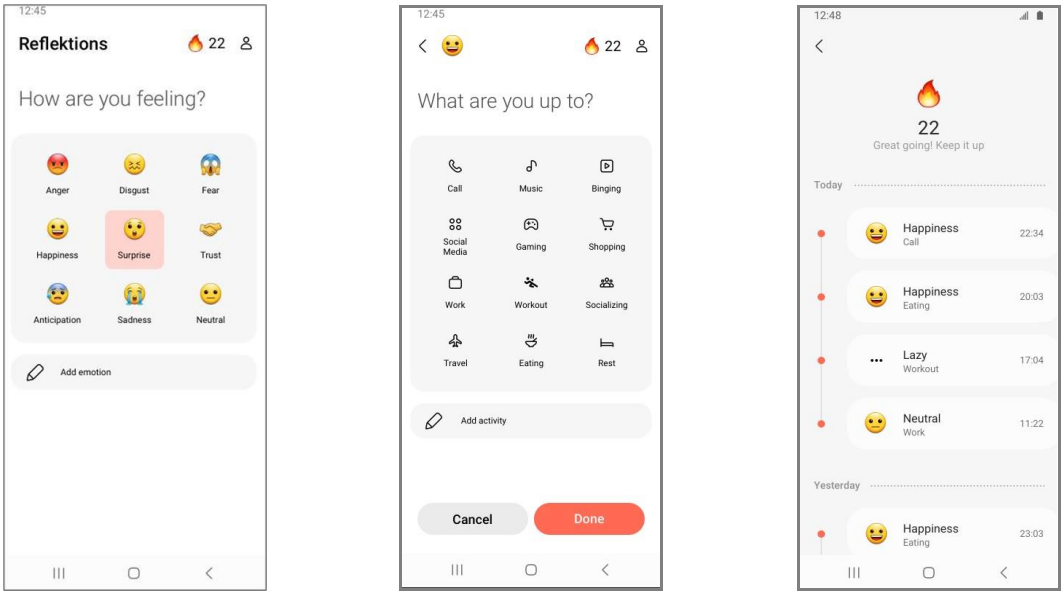
More interesting approaches to collect labelled data using smartphones have been explored in the following years. MoodExplorer (Zhang et al., 2018) proposes the system framework for compound emotion detection via smartphone sensing. They collect data from 30 university students and apply feature selection techniques to detect compound emotions rather than singular labels. Morshed et al. (2019) propose a framework for predicting mood instability using passively sensed data gathered from smartphones and wearables. Authors formulated mood prediction as regression problem and considered several modalities based on audio, sensor, GPS information etc. to predict mood instability. Darvariu et al. (2020) developed MyMood, a smartphone application that allows users to periodically log their emotional state together with pictures from their everyday lives along with passive sensor data. They use this visual information along with sensor data to develop deep learning methods for human emotion prediction. Roshanaei et al. (2017) developed the Emosensing app to collect ground truth data for emotion prediction, which requires the user to launch the application and log the emotion manually. They incentivize users with monetary benefits to encourage data logging. Further, they used the collected data: activity, smartphone app usage, and location (private data) to predict 13 different user emotions. Further studies also use personality information and relate to users' emotional states, as studied in (Donovan et al., 2021).

Most of these prior works rely on the content of user activity, like message content which may contain user-private information. Further, most studies collect data by incentivizing the user to record emotions which may result in unnatural and forced data. We address these drawbacks in our work and propose a non-intrusive smartphone activity annotation technique followed by some significant insights.

3 Designing Emotion Data Collection Experience

Designing the most effective application for data collection is a challenge. We try to motivate the user to regularly contribute annotations without the need to provide any incentive for the same. Doing this will help collect accurate data annotation and avoid forced data logging. To achieve this, we develop a novel design for data collection, which is discussed in this section.

We design our emotion annotation method following an iterative and user-centric approach. We designed the experience not only to enable users to record emotion-related data with ease but also to reflect on the data recorded insightfully. We tried to encounter the main problem, which any data recording app might face (Caldeira et al., 2017), which is to motivate or trigger users to record the data every day in a regular fashion. We solved this by studying social media behaviors where users feel rewarded subconsciously for using the app, e.g., Gamification in Snapchat (Hamari et al., 2014). They introduced a feature called Streaks that encourages users to open the app and use it regularly (Furner et al., 2014) We took the inspiration from the same and incorporated Streaks in our design to motivate users for frequent data annotation as shown in Figure 1c.



a. Emotion selection screen b. Activity selection screen c. Streak page

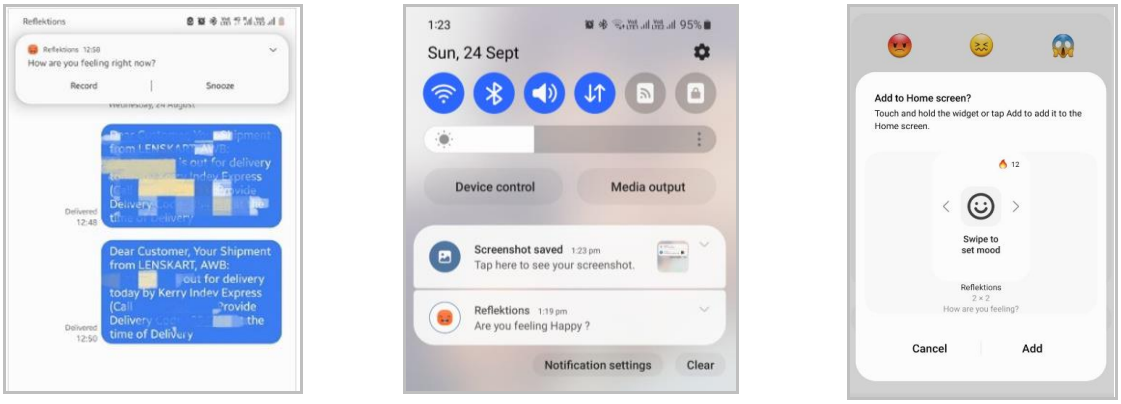
Figure 1. Main Screens of the Reflektion Application.

We anticipated that users might find it challenging to find the app repeatedly, as they might not have placed it in phone in an accessible manner (Lavid Ben Lulu et al., 2013). We thus introduced a nudge, which prompts the user to add the app widget on the home screen for quick access. This nudge surfaces during the first-time experience of the app (Figure 2c). Once the user adds the widget in the home screen, they can record emotion with a simple tap.

We added periodic notifications, which serve as a reminder for the user to log their emotion. Apart from these nudges, we also trigger notifications after certain activities like a switch in airplane mode, phone reboot, playing music, etc., which allow us to collect data in that particular context. We believe that users can then reflect on the emotion they viewed just after the event/ situation has ended (Ferdinando et al., 2018).

We used basic design guidelines, simple language (ex: what are you feeling), comprehensive symbols (ex: emoticons for conveying emotions), for application's ease of use, quick learnability, and logging repeated and accurate input. Before the participants start providing emotion data, our data collection Reflektion app requires one time on-boarding setup. During on-boarding, participants are asked to provide basic demographic data (age-range, gender and occupation status) and big five personality traits (extraversion, openness, conscientiousness, agreeableness, and neuroticism). Each trait is provided with three descriptors as low, medium and high which represent level of each personality trait. To ensure most appropriate selection for personality traits by participants, app UI provides description and expected behaviour for each trait. Users can then easily follow the simple three-step

process to reach the emotion selection page through various nudges, select emotion, and select related activity to record the emotion data. We initially started with nine emotion categories and further expanded to three more emotion categories based on user feedback in subsequent release of the application. In the future, we aim to provide insights based on the patterns emerging from the emotion data on the application screen.



a. Silent Notifications b. Revisiting Notifications c. Widget Nudge

Figure 2. Periodic notifications for data collection & Widget nudge for Data annotation.

4 System Design and Data Collection

Post designing the user interface and application structure, implementing an effective system functionality is another challenge. We take inspiration from LiKamWa et al (2013) and attempt to further eliminate some gaps. We observe a need to collect the smartphone activity data in the background before and after the user records their emotion to capture the details of what could have led to the emotion and the user behavior in a particular emotional state. Our data collection window lasts 45 minutes, and we trigger a notification 15 minutes into the collection window (data collected for 15 minutes before user-annotation, 30 minutes post the annotation). For user convenience, we prompt the user to annotate the data with a frequency of 4 times daily, between 9am and 9 pm. This ensures that participants do not disable the app, fearing battery drain. However, the participants can enter emotion anytime by directly launching the application. Participants were made aware of the data and data collection process to ensure transparency. We design the data collection, and annotation application with these policies after detailed user-trial experimentation and research (LiKamWa et al., 2013). For deciding the data collection window, we take cues from previous studies by MoodExplorer (Zhang et al., 2017) and EmoSensing (Roshanaei et al., 2017) and experimented with one-hour window size during the early phase of data collection. However, we observe that multiple emotions are being reported by participants due to the large window size, leading to undesired data collection. So, based on user feedback and our data analysis, we reduced the data collection window to 45 minutes.

The designed application allows the user to record emotion in multiple ways: Through Scheduled Notifications, the widget, and self-initiated, as explained in the previous section. The application system design has mainly three core components shown in the Figure 3: 1) Work Manager, 2) Monitors, and 3) Logger.

The Work Manager is responsible for scheduling tasks periodically by transitioning among different states. These states are None, Monitor, Notification, and Cleanup. The application enters the None state immediately after installation, then enters the Monitor state where we register for required Monitors to capture data. We then register for the Notification state with a 15-minute delay. In this case, we give the user a gentle nudge to enter their current emotion, with an optional activity entry and a short note that allows us to capture the trigger for the entered emotion. The monitors continue to

function while in the notification state. The work manager registers for the cleanup state with a 30-minute delay after the notification, where we unregister all monitors and retrieve the collected data from the database to convert it to JavaScript Object Notation (JSON) files. The Monitor state is registered with 120-minute delay. The Logger then attempts to log the collected data into an Amazon Web Service (AWS) Simple Storage Service (S3) bucket that holds the data accumulated from all the users. We determined the duration of these delays through extensive user-trial experimentation.

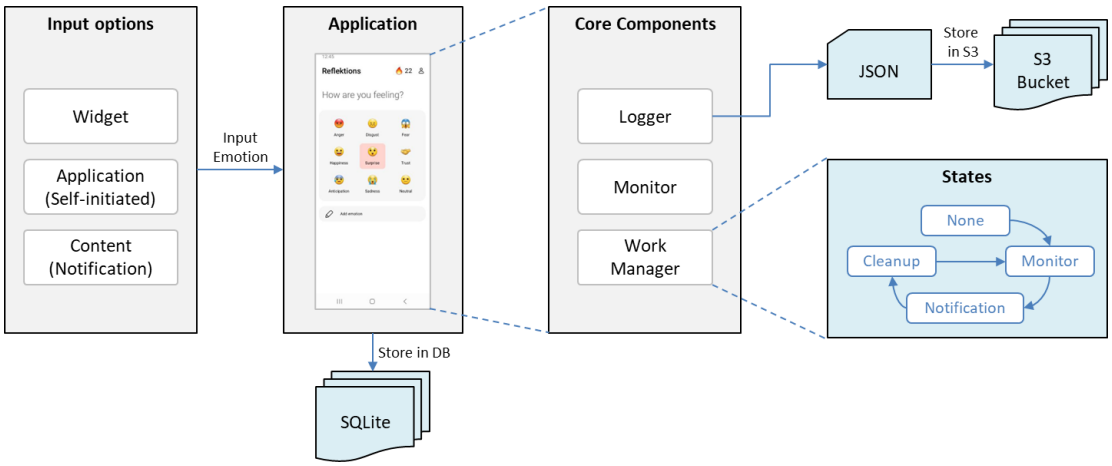


Figure 3. Periodic notifications for data collection & Widget nudge for Data annotation.

The Monitors are various broadcast receivers and listeners that help capture significant user activities like App usage time, Wi-Fi connection/disconnection, Flight Mode On/Off, etc. Some features like Call duration are further used to derive essential elements like average call duration for the top 10 most contacted individuals. These are non-intrusive features collected through user consent. We ensure to mask any PII if present in any collected user data.

The Logger takes care of sending the user data to the AWS S3 Bucket. This first checks for any working network connection. If the device is connected, it checks for any pending JSON files for logging. Once we determine that there are indeed files pending, the Logger fetches the JSON files and attempts to log into the server. Once successful, the files are deleted from internal memory. Note that we use the secure android identifier to uniquely identify each device and avoid any personal information from being acquired. The files corresponding to respective devices are stored in separate folders.

5 Emotional Data Analysis

One of the crucial components of the entire process is data analysis. Analytics is used to process the collected raw data and further turn it into insights (that can unearth new relationships between features). Additionally, we can use these insights to provide actionable recommendations and feature improvement suggestions as shown in Figure 4. This section provides an example of different analytics performed on the collected data, to demonstrate its potential.

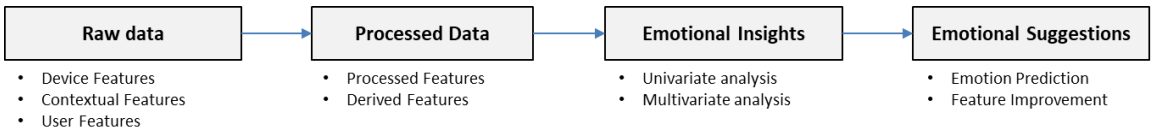


Figure 4. Data Processing pipeline.

5.1 Research context

The android-based annotation application called Reflektion is used to collect the data. The annotated data is collected from November 2022 to April 2023. As mentioned in previous sections, this study aims to understand user conduct, emotion, and personality and explore relationships that can enable novel use cases. The participants used Android-based smartphones having Android Pie and above version. The participants are given the freedom to log the emotion data or ignore the prompt, thereby eliminating any need for forced entry to strengthen the authenticity of the data further.

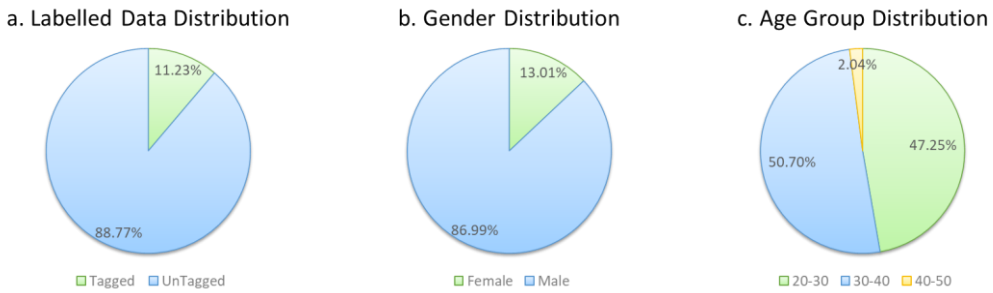


Figure 5. Ground Truth Labels Tagging & Demographic Data Distribution

As shown in Figure 5a, approximately 11 percent of the data collected is tagged by the user with emotion. 13% of participants are female and 87% male. A major percentage (98%) of participants are within the age group of 20-40 years. This demographic distribution is covered in Figure 5b and 5c respectively. This dataset is not entirely representative of the user's smartphone usage behavior, mainly due to demographic constraints (all the participants are from a specific country). Also, all the participants are employed. Nevertheless, the outcomes highlight the approach's unique value.

5.2 Data visualization and emotional insights

Under this, we plot individual features and try to derive insights from the same. One crucial finding is that the existing eight basic emotions are not enough to capture user emotions non-intrusively. Instead, few other emotion categories get priority over the different categories of eight basic emotions, as shown in Figure 6a. Figure 6b shows the distribution of other emotions manually entered by the participants.

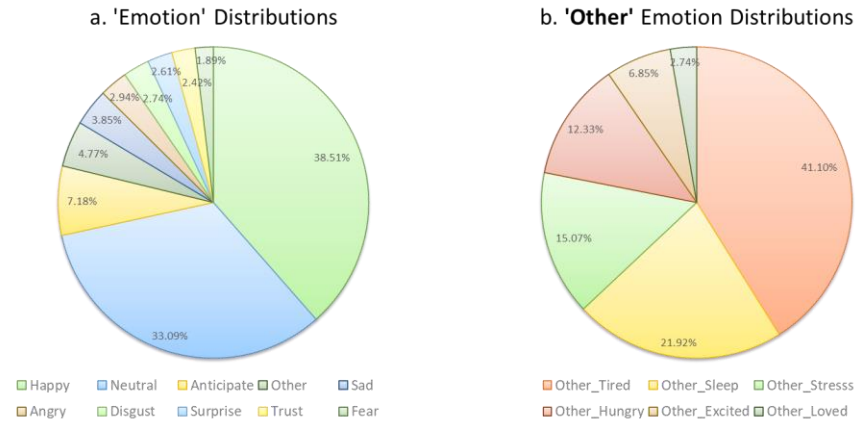


Figure 6. Emotion Distribution labelled by the Users.

The distribution of significant application categories usage for various emotions is covered in Figure 7. A social category app is used mainly by the user in a happy or neutral emotional state. One fascinating insight is that user tends to use entertainment applications when they are sad. When the user is disgusted, music-related applications are rarely used. Apart from happy and neutral emotions, when the user's emotional state is that of trust as well, music applications are used.

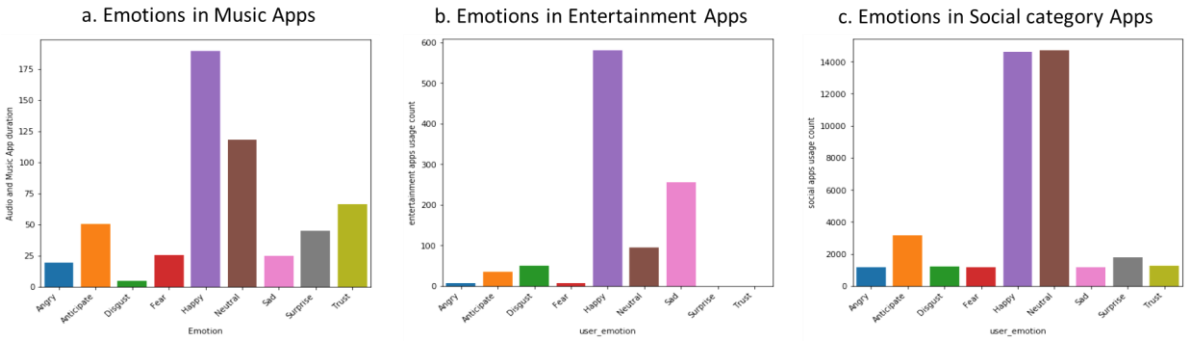


Figure 7. Emotion Distribution in different category of applications (App Usage vs Emotion).

Users tend to lock/unlock phones more when in an anticipation state as shown in Figure 8a. In anticipation state, apart from high lock/unlock, foreground time (refer Figure 8c) and food/drink category app usage is also high compared to other emotional states. This can be when user orders food, keeps lock/unlock the phone and enters into anticipation state. Figure 8a shows lesser lock/unlock frequency for sad emotion state and Figure 8c also shows comparatively low values for foreground time for sad emotion. From these observations, it can be weakly inferred that user tends to use phone less when in sad state. High incoming calls are associated with anticipation state can be seen in Figure 8b. For all the graphs in Figure 8, X-axis represents count/usage (in minutes) of feature fitted in fixed size bins, Y-Axis represents number of Data points available in the bins.

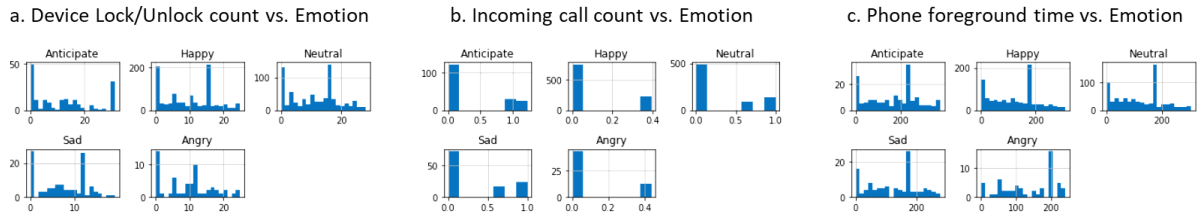


Figure 8. Emotion Distribution among various smart phone features.

Figure 9 captures emotion dynamics of the emotion data logged by the user. Figure 9a captures the count of different emotions entered by individual users over the data collection period. Figure 9b captures the distribution of emotion entered by various users in sequential manner affirming the fact that user undergoes through various emotions over a period of time and our data captures the variation.

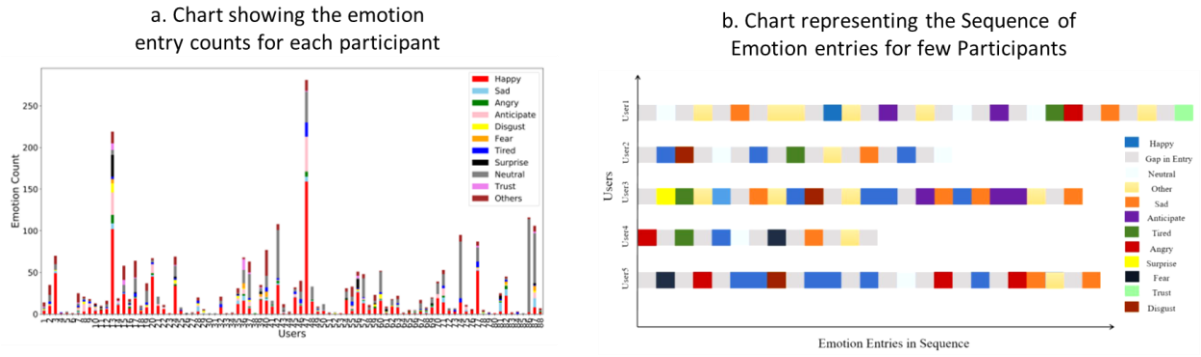


Figure 9. User Emotion Dynamics.

6 Feature Engineering

As a first step, we collect user data through android based Reflektion App as mentioned in section 4. Further, this data is encrypted and sent to the server in JSON format. JSON data files are downloaded and further processed to convert into Comma Separated Values (CSV) format in the next step to draw actionable insights and model development. We use python-based environment for all our data pre-processing and model development.

Figure 10 describes our proposed model pipeline. We use heat map-based feature correlation analysis to understand the correlation between features. We identify that 'application added' and 'application replaced' are highly correlated and hence dropped the 'application replaced' feature. We also observe that some of the captured features always have null values such as physical activity corresponding to foot and tilting, and application-changed events, so we decided to drop these.

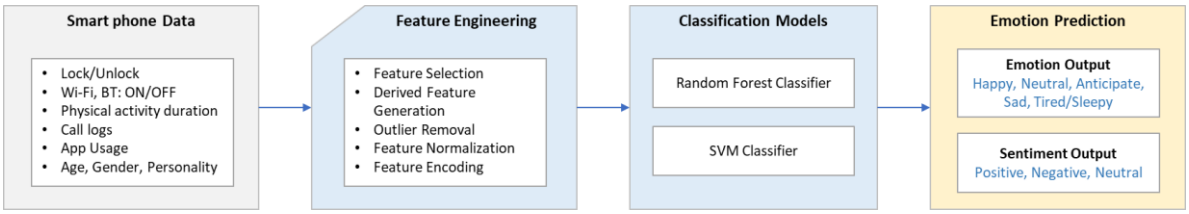


Figure 10. Overview of Model Prediction Pipeline, with Novel smartphone data.

As a part of the pre-processing pipeline, we first remove outliers from non-categorical features and then apply standard scaler normalization on these features. For identifying outliers, we use the inter quantile range (IQR) to identify and replace outliers by mean values of the respective features. In our experiment, replacing outliers by mean instead of Q3 (third quantile) gives better results. We then encode categorical features such as age group, user activity, time of the day, five personalities (openness, agreeable, conscientiousness, extraversion, neuroticism), and phone mode using one hot encoding method and use label encoding for our target variable 'user emotion' and gender. For getting the time of the day, the entire day is divided into four segments: Morning (6-12), Noon (12-16), Evening (16-20), and Night (20-6) which makes it a categorical feature for our analysis. After the feature encoding step, we get a total of 59 features including the target variable which we use for model training and validation. As discussed in section 5.2, we get several emotion labels tagged by users but for the emotion prediction task, we drop those emotions whose frequency is less than 2% across the data. Consequently, we get five emotions viz Happy, Neutral, Anticipate, Tired/Sleepy, and Sad. We drop other emotions as the number of data points is not sufficient for training and validation. As shown in Data Visualization and Insights, the dataset is highly imbalanced with emotions such as happy and neutral having a higher proportion of samples compared to the other emotions. So, to

handle the data imbalance issue, we try different standard sampling approaches such as Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) and SMOTE-based over and under-sampling strategies as well as class weights for better generalization of the model on rarer classes. In our experiments, we get better results using class weights compared to other methods. So, for all experiments, we adopt class weights as a standard method for handling class imbalance.

Apart from emotion prediction, we also provide predictions for three sentiments polarity viz. positive, negative, and neutral on the same dataset. To generate sentiment ground truth for the data points, we use a study by Goncalves et al (2017) to group different emotions into sentiments. Unlike for emotion task where emotions having frequency lesser than 2% are dropped, we consider all emotions for getting sentiment labels except the surprise and anticipation emotions as they can be both positive and negative based on the user activity outcome.

7 Emotion Prediction: Result & Analysis

In this section, we discuss the results related to our annotation and emotion prediction methods. One significant point is that our method does not incur a higher mental workload than filling annotations using the widely-used Self-Assessment Manikin (SAM) method (Margaret et al.).

We perform a detailed comparative study between prior works involving smartphone-based annotation techniques to perform various emotion prediction tasks as shown in Table 2. We observe that the Reflektions application fairs better in terms of preserving user privacy and capturing natural emotion as no private data is used and user smartphone-related tasks are not hampered. Additionally, we also make use of demographic and personality information given by the user to improve prediction results which is not explored by prior works. We are also able to collect sizeable annotated data samples without any monetary rewards for the participants.

Features (O: Yes, X: No)	**MoodExplorer**	**iSelf**	**MoodScope**	**EmoSensing**	**Reflektion (Proposed)**
Collection methodology	Daily user tasks	Daily user tasks	Field study group	Recruited users	**Daily user tasks**
User feature (age, sex, etc.)	X	X	X	O	**O**
Use of personality information	X	X	X	X	**O**
Device Features (Bluetooth, Lock/Unlock, etc.)	O	O	O	X	**O**
User App Usage Behavior (Duration, category etc.)	O	O	O	O	**O**
Private Content (SMS, Location, Images, etc.)	O	O	O	O	**X (Private content not used)**
Monetary Rewards for participants	O	X	O	O	**X**
Number of Participants	30	10	32	27	**100**
Number of Data samples	X	3600	X	X	**13700**
Number of Annotated samples	X	3600	X	X	**1532**
Task Type	Compound Emotion	Emotion	Mood	Multiclass Emotion	**Emotion, Sentiment**

Table 1. Comparative study between various smartphone-based annotation applications.

Since there is no public dataset available for benchmarking emotion and sentiment prediction using mobile phone usage data, we could not provide a comparison of our method performance with respect to State of the art (SOTA) and instead validated our dataset using standard models. For our labelled dataset, we train both machine learning (ML) and deep learning (DL) models to classify emotions and sentiments. Since our labelled dataset is small, ML models are found to be more effective than DL models. As shown in Table 3, we train several tree-based models such as Random Forest (Breiman, 2001), Support Vector Machine (SVM) – Radial Basic Function (RBF) Kernel (Hearst et al., 1998), XGBoostClassifier (Chen et al., 2016), Gradient Boosting Classifier (Chen et al., 2016), CatBoost (Dorogush et al., 2018) and Light Gradient Boosting Machine (LightGBM) (Guolin et al., 2017) for our tasks. We divide the dataset into 80:20 training and testing ratio, with 20% of the data points used for testing. For every trained model, we apply grid search-based hyperparameter tuning and observe the performance improvement in nearly all the models. For fine tuning tree-based models, mostly two parameters are considered: max depth and n estimators which represent the depth of the tree and the

number of trees respectively. As shown in Table 3, we get the best performance for the emotion prediction task using Random Forest with the following hyperparameters: max depth=18; n estimators=300; min samples leaf =5. Similarly, for the sentiment task, SVM with hyperparameters as gamma=0.009; kernel = 'rbf'; C =3 provide the best performance.

For comparison, we also train two deep learning models namely Multi-Layer Perceptron (MLP) and 1D convolution. MLP consists of three dense layers with units viz 32, 64, and 128 in sequence. Similarly, the 1D convolution model consists of three 1DConv layers having kernel sizes of 32, 64, and 128 followed by a 128-dense layer. The last layer for both the models is the classification layer with 5 units for the emotion task and 3 units for the sentiment task. For both models, we use the activation function as Rectified Linear Unit (ReLU) for intermediate layers and SoftMax for the classification layer. For hyperparameter tuning and to get an optimum number of layers for both models, we use the Keras Tuner framework, and above discussed configuration provided the best performance. Other hyperparameters are as follows: learning rate = .0001; batch size = 64; epochs = 100.

Model	Emotion Prediction			Sentiment Prediction		
	Accuracy (%)	F1 Score (%) (Macro Average)	AP Score (Micro)	Accuracy (%)	F1 Score (%) (Macro Average)	AP Score (Micro)
SVM (RBF)	65.21	53	**.73**	**77.95**	**78**	.82
XGBClassifier	51.71	43	.48	72.60	71	.81
Gradient Boosting Classifier	49.42	39	.46	73.05	73	.80
CatBoost	66.59	55	.70	73.49	72	.83
LGBClassifier	52.40	45	.51	73.49	73	.82
Random Forest	**67.73**	**56**	.72	71.26	71	.80
MLP (3 Layer)	62.92	52	.64	73.49	74	.82
1D Convolution (4 Layer)	61.09	48	.70	71.26	67	**.84**

Table 2. Performance of different classifiers on Emotion and Sentiment prediction tasks.

Since our training and validation data is highly imbalanced (Happy: 45%, Neutral: 33%, Anticipate: 8%, Sad: 5%, Tired/Sleepy: 9%), we calculate F1 score, Average Precision (AP) for both the sentiment (3 classes: positive, negative, and neutral) and emotion (5 classes: Anticipate, Happy, Neutral, Sad and Tired/Sleepy) prediction tasks. F1 score (macro average) is a simple average over all classes, so each class is given equal weight independent of their proportion and AP (Micro) takes into account the class imbalance in calculating average precision. Table 3 shows the performance of several models for both tasks. For the emotion prediction task, the Random Forest algorithm provides the best performance among others achieving 67.73% accuracy, 56 % F1-score, and an AP score of .72. The SVM (RBF kernel) outperformed other models with an accuracy of 77.95%, F1-score of 78% on the sentiment prediction task.

Metrics	Emotion Prediction (Random Forest)					Sentiment Prediction (SVM)		
	Anticipate	Happy	Neutral	Sad	Tired/Sleepy	Negative	Neutral	Positive
Precision	49	70	76	55	53	80	70	83
Recall	59	87	56	26	46	85	72	79
F1 Score	53	78	64	35	49	82	71	81

Table 3. Performance results for Emotion and Sentiment prediction tasks.

Table 4 provides class-wise precision, recall, and F1-score by best performing Random Forest and SVM models on emotion and sentiment prediction tasks respectively. Low scores for Sad and Tired/Sleepy emotions can be attributed to the lower number of samples in the data distribution.

We perform an ablation study to analyse and understand the impact of demography (age, gender) and personality (five personality traits) as input features to the emotion model. Table 5 discuss experimentation performed for top-performing ML models for the emotion prediction task. It can be seen that there is always performance improvement whenever either demography or personality or both are taken as input to the model. Another interesting observation is, when personality alone is considered as input (dropping demography features) along with other features, we get the best performance for SVM and fair performance on the Random Forest model. It establishes the fact that there is implicit relation between emotion and personality, and knowing the user's personality can be beneficial in determining the user's emotional state. We also observed a little performance drop when both demography and personality features are considered. One possible reason can be high data imbalance with respect to gender and age group (refer Figure 5) leading to downward performance.

Model	**Demography**	**Personality**	**F1 Score (%)** (Macro Average)	**AP Score** (Micro)
SVM (RBF)	X	X	51	.72
	O	X	53	.73
	X	**O**	**54**	**.74**
	O	O	53	.73
Random Forest	X	X	48	.64
	O	X	49	.65
	X	O	55	.71
	O	**O**	**56**	**.72**

Table 4. Ablation Study for Emotion Task.

We notice subpar performance for the emotion prediction task in comparison to the sentiment prediction task. It is particularly due to 1) the relatively lesser number of samples for the classes such as Anticipate, Sad, and Tired/Sleepy. 2) feature set for rare classes like Sad, Tired/Sleepy, etc., is highly overlapping with classes such as happy and neutral as demonstrated in the t-SNE plot in Figure 11a and thus making it challenging for the models to learn exact boundaries. Relatively better performance for the sentiment prediction task can be attributed to the reduced complexity of the task as a result of grouping emotions into sentiments as confirmed by the t-SNE plot in Figure 11b. The t-SNE plot shows that both positive and neutral class data points form separate clusters (except for a few outliers) while negative class data points lie mostly in the periphery of the other two classes making the model learn the pattern and perform better.

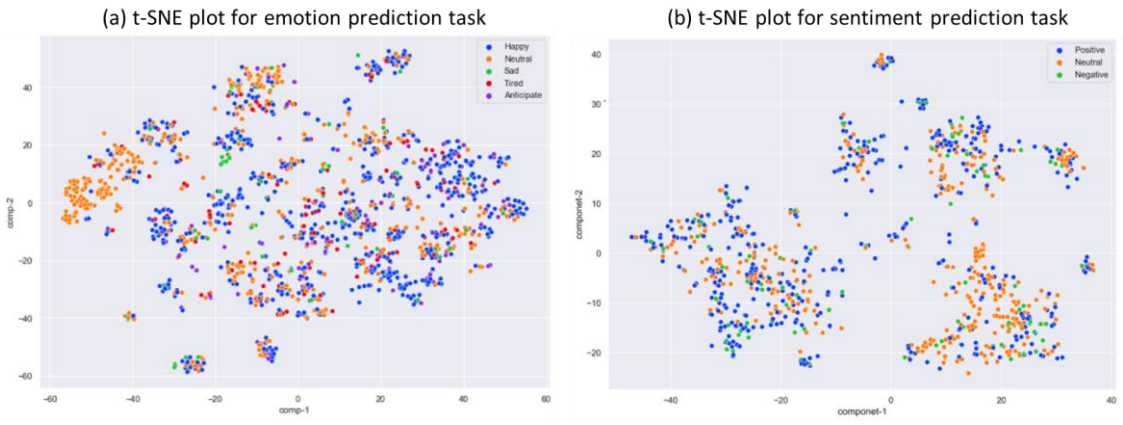


Figure 11. t-SNE plot for emotion and sentiment prediction tasks using categorical features.

In Figure 12, we demonstrate various performance metrics for Random Forest on emotion prediction task. Figure 12a shows the True Positive Rate (TPR) versus False Positive Rate (FPR) graph, called as Area Under the Curve-Receiver Operating Characteristic (AUC-ROC) curve, across varying thresholds for target emotion classes. The ROC curve with better discrimination ability lies top left corner. It can be seen that Neutral and Anticipate have greater discriminate abilities than Sad. Figure 12b shows class-wise Precision-Recall (PR) performance (referred as class-wise AUC-PR curve). For each class, it is calculated by considering the problem statement as binary and represents the discrimination capacity of individual class versus others. In our case, we perform target class vs others analysis to get class wise average precision score. It can be seen that Happy and Neutral perform comparatively better and Sad has the lowest precision score. Figure 12c shows a trade-off between precision and recall for our emotion prediction task, micro-averaged over all the classes.

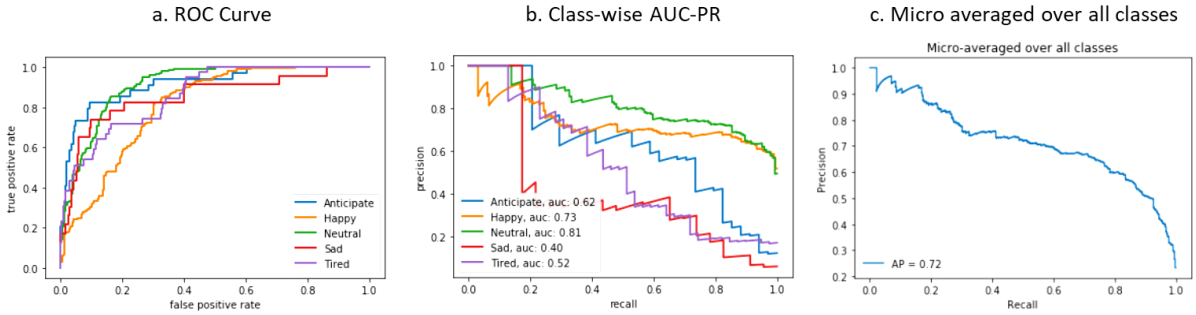


Figure 12. Performance of Random Forest classifier on Emotion Prediction task.

Given the challenges of designing data collection applications for smartphones, there are a few limitations to our work. First, the Reflektion app is designed to capture real-world smartphone interactions non-intrusively and strictly adhere to user privacy. This prohibits us to use user content like chat messages, pictures etc. which if utilized can naturally increase the accuracy. Still, we believe our findings provide a first step towards collecting more precise emotional ground-truth labels without compromising user privacy. Second, while we designed and iterated alternatives for inputting real-time emotion annotation, we explored external factors/activities influencing emotion in a limited way. However, our aim here was to firstly validate how well our designed annotation method works in comparison to the standard practice of asking participants to log emotions specifically. Further, we also observed imbalance in our dataset (Gender and age-group, emotion data distribution), which negatively impacted model predictions. However, we overcome these limitations during the preprocessing phase by using appropriate feature engineering strategies.

8 Conclusion and Future Works

We presented a design for a real-time emotion annotation technique for smartphone usage behavior without invading user privacy. Our approach enables researchers to collect emotion annotations while using a smartphone. Through collection methodology, we ensure that our method does not incur extra pressure to log emotions, as emotion entry is optional. This also ensures that the logged data capture genuine emotions. Moreover, we verified the consistency of annotations by deriving new insights about user behavior and strengthening insights derived from previous work in the field. We further use the annotated data to predict user emotions and sentiment with good accuracy. We also demonstrated that using the user personality as one of the features, emotion prediction accuracy can be enhanced. Once user emotion is known, one can enable various novel on-device and multi-device experiences. It is possible to achieve deeper personalization in applications like Music players, Search, Contacts, etc. on the device where recommendations can take user emotion into account. Our work underscores the importance of collecting ground truth emotion annotations, which is essential for ensuring accurate

user emotion recognition non-intrusively. Further, in the future, we plan to explore deep learning methodologies like zero-shot learning to increase the accuracy of emotion prediction. Also, we intend to distribute a mobile application across geographies and a few educational institutions to improve the demographic distribution of data and to achieve better generalization.

References

Akhand, M.A.H., Roy, S., Siddique, N., Kamal, M.A.S. and Shimamura, T., (2021). "Facial emotion recognition using transfer learning in the deep CNN." *Electronics*, *10*(9), p.1036.

Morshed, M.B., Saha, K., Li, R., D'Mello, S.K., De Choudhury, M., Abowd, G.D. and Plötz, T., (2019). "Prediction of mood instability with passive sensing." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(3), pp.1-21.

Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F. and Pentland, A., (2014, November). "Daily stress recognition from mobile phone data, weather conditions and individual traits." In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 477-486).

Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S., (2008). "IEMOCAP: Interactive emotional dyadic motion capture database." *Language resources and evaluation*, *42*, pp.335-359.

Caldeira, C., Chen, Y., Chan, L., Pham, V., Chen, Y. and Zheng, K., (2017). "Mobile apps for mood tracking: an analysis of features and user reviews." In *AMIA Annual Symposium Proceedings* (Vol. 2017, p. 495). American Medical Informatics Association.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., (2002). "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research*, *16*, pp.321-357.

Chen, T. and Guestrin, C., (2016, August). "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Darvariu, V.A., Convertino, L., Mehrotra, A. and Musolesi, M., (2020). "Quantifying the relationships between everyday objects and emotional states through deep learning based image analysis using smartphones." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(1), pp.1-21.

data.ai (2022). *The State of Mobile in 2022: How to Succeed in a Mobile-First World As Consumers Spend 3.8 Trillion Hours on Mobile Devices.* Available at: https://www.data.ai/en/insights/market data/state-of-mobile-2022/.

Dixon, T., (2012). "Emotion: The history of a keyword in crisis." *Emotion Review*, *4*(4), pp.338-344.

Donovan, R., Johnson, A., deRoiste, A. and O'Reilly, R., (2021). "A Multimodal Workflow for Modeling Personality and Emotions to Enable User Profiling and Personalisation." In *VISIGRAPP (2: HUCAPP)* (pp. 145-152).

Dorogush, A.V., Ershov, V. and Gulin, A., (2018). "CatBoost: gradient boosting with categorical features support." *arXiv preprint arXiv:1810.11363*.

Ekman, P., (1992). "An argument for basic emotions." *Cognition & emotion*, *6*(3-4), pp.169-200.

Ferdinando, H. and Alasaarela, E., (2018, January). "Enhancement of emotion recogniton using feature fusion and the neighborhood components analysis." In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*. SCITEPRESS Science And Technology Publications.

Hoefel F Francis T (2022). *Generation Z and its implications for companies.* Available at: https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/true-gen-generation-z-and-its-implications-for-companies.

R Fulcher (2022). *Create intuitive and beautiful products with Material Design*. Available at: https://material.io/design.

Furner, C.P., Racherla, P. and Babb, J.S., (2014). "Mobile app stickiness (MASS) and mobile interactivity: a conceptual model." *The Marketing Review*, *14*(2), pp.163-188.

Ghosh, S., Hiware, K., Ganguly, N., Mitra, B. and De, P., (2019). "Emotion detection from touch interactions during text entry on smartphones." *International Journal of Human-Computer Studies*, *130*, pp.47-57.

Gonçalves, V.P., Giancristofaro, G.T., Filho, G.P., Johnson, T., Carvalho, V., Pessin, G., Neris, V.P.D.A. and Ueyama, J., (2017). "Assessing users' emotion at interaction time: a multimodal approach with multiple sensors." *Soft Computing*, *21*, pp.5309-5323.

Smith, D.P., Keyfitz, N. and Graunt, J., (1977). "Natural and political observations mentioned in a following index, and made upon the bills of mortality." *Mathematical Demography: Selected Papers*, pp.11-20.

Guenther, P.M., Cleveland, L.E., Ingwersen, L.A. and Berline, M., (1994). "Questionnaire development and data collection procedures." *Design and operation: the continuing survey of food intakes by individuals and the Diet and Health Knowledge Survey*, *96*, pp.42-63.

Habitics (2022). *Daylio Journal - Mood Tracker.* Available at: https://play.google.com/store/apps/details?id=net.daylio&hl=en&gl=US&pli=1.

Hamari, J., Koivisto, J. and Sarsa, H., (2014, January). "Does gamification work?--a literature review of empirical studies on gamification." In *2014 47th Hawaii international conference on system sciences* (pp. 3025-3034). Ieee.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B., (1998). "Support vector machines." *IEEE Intelligent Systems and their applications*, *13*(4), pp.18-28.

Hung, G.C.L., Yang, P.C., Chang, C.C., Chiang, J.H. and Chen, Y.Y., (2016). "Predicting negative emotions based on mobile phone usage patterns: an exploratory study." *JMIR research protocols*, *5*(3), p.e5551.

Inexika Inc. (2022). *IMoodJournal*. Available at: https://play.google.com/store/apps/details?id=com.inexika.imood.

John, O.P. and Srivastava, S., (1999). "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives."

Kanjo, E., Younis, E.M. and Ang, C.S., (2019). "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection." *Information Fusion*, *49*, pp.46-56.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., (2017). "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in neural information processing systems*, *30*.

Kołakowska, A., Szwoch, W. and Szwoch, M., (2020). "A review of emotion recognition methods based on data acquired via smartphone sensors." *Sensors*, *20*(21), p.6367.

Lavid Ben Lulu, D. and Kuflik, T., (2013, March). "Functionality-based clustering using short textual description: Helping users to find apps installed on their mobile device." In *Proceedings of the 2013 international conference on Intelligent user interfaces* (pp. 297-306).

LiKamWa, R., Liu, Y., Lane, N.D. and Zhong, L., (2013, June). "Moodscope: Building a mood sensor from smartphone usage patterns." In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services* (pp. 389-402).

Little, B., (2014). *Me, myself, and us: The science of personality and the art of well-being*. Public Affairs.

Livingstone, S.R. and Russo, F.A., (2018). "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." *PloS one*, *13*(5), p.e0196391.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A. and Cambria, E., (2019, July). "Dialoguernn: An attentive rnn for emotion detection in conversations." In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 6818-6825).

Marshall, G., (2005). "The purpose, design and administration of a questionnaire for data collection." *Radiography*, *11*(2), pp.131-136.

McAdams, D.P., (2015). *The art and science of personality development*. Guilford Publications.

McCrae, R.R. and Costa, P.T., (2008). "Empirical and theoretical status of the five-factor model of personality traits." *The SAGE handbook of personality theory and assessment*, *1*, pp.273-294.

Mirjafari, S., Masaba, K., Grover, T., Wang, W., Audia, P., Campbell, A.T., Chawla, N.V., Swain, V.D., Choudhury, M.D., Dey, A.K. and D'Mello, S.K., (2019). "Differentiating higher and lower job performers in the workplace using mobile sensing." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(2), pp.1-24.

Montag, C., Błaszkiewicz, K., Sariyska, R., Lachmann, B., Andone, I., Trendafilov, B., Eibes, M. and Markowetz, A., (2015). "Smartphone usage in the 21st century: who is active on WhatsApp?." *BMC research notes*, *8*(1), pp.1-6.

Ooi, C.S., Seng, K.P., Ang, L.M. and Chew, L.W., (2014). "A new approach of audio emotion recognition." *Expert systems with applications*, *41*(13), pp.5858-5869.

Pelteret, M. and Ophoff, J., (2016). "A review of information privacy and its importance to consumers and organizations." *Informing Science*, *19*, pp.277-301.

Plutchik, R., (2001). "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice." *American scientist*, *89*(4), pp.344-350.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E. and Mihalcea, R., (2018). "Meld: A multimodal multi-party dataset for emotion recognition in conversations." *arXiv preprint arXiv:1810.02508*.

Revelle, W. and Scherer, K.R., (2009). "Personality and emotion." *Oxford companion to emotion and the affective sciences*, *1*, pp.304-306.

Roshanaei, M., Han, R. and Mishra, S., (2017, July). "Emotionsensing: Predicting mobile user emotion." In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 325-330).

Russell, J.A. and Mehrabian, A., (1977). "Evidence for a three-factor theory of emotions." *Journal of research in Personality*, *11*(3), pp.273-294.

Sadeghian, A. and Kaedi, M., (2021). "Happiness recognition from smartphone usage data considering users' estimated personality traits." *Pervasive and Mobile Computing*, *73*, p.101389.

Schobel, J., Schickler, M., Pryss, R., Maier, F. and Reichert, M., (2014). "Towards process-driven mobile data collection applications: Requirements, challenges, lessons learned."

Breiman, L., (2001). "Random forests." *Machine learning*, *45*, pp.5-32.

Sun, B., Ma, Q., Zhang, S., Liu, K. and Liu, Y., (2017). "iSelf: Towards cold-start emotion labeling using transfer learning with smartphones." *ACM Transactions on Sensor Networks (TOSN)*, *13*(4), pp.1-22.

Talevich, J.R., Read, S.J., Walsh, D.A., Iyer, R. and Chopra, G., (2017). "Toward a comprehensive taxonomy of human motives." *PloS one*, *12*(2), p.e0172279.

Thriveport, LLC (2021). *MoodKit*. Available at: http://www.thriveport.com/.

Track Your Happiness.org (2022). *Track Your Happiness*. Available at: https://go.trackyourhappiness.org/.

Wakefield, J., (2022). "People devote third of waking time to mobile apps." *BBC. January*, *12*.

Wang, Y., Zhang, J., Ma, J., Wang, S. and Xiao, J., (2020, July). "Contextualized emotion recognition in conversation as sequence tagging." In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue* (pp. 186-195).

You, Q., Luo, J., Jin, H. and Yang, J., (2016, February). "Building a large scale dataset for image emotion recognition: The fine print and the benchmark." In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 30, No. 1).

Ramos, F.Y., (2014). "Not all emoticons are created equal." *Linguagem em (Dis) curso*, *14*(3), p.5.

Zhang, T., El Ali, A., Wang, C., Hanjalic, A. and Cesar, P., (2020, April). "Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

Zhang, X., Li, W., Chen, X. and Lu, S., (2018). "Moodexplorer: Towards compound emotion detection via smartphone sensing." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(4), pp.1-30.