

Chapter 9

Affective and Personality Corpora

Ante Odić, Andrej Košir and Marko Tkalčič

Abstract In this chapter we describe publicly available datasets with personality and affective parameters relevant to the research questions covered by this book. We briefly describe the available data, acquisition procedure, and other relevant details of these datasets. There are three datasets acquired through the users' natural interaction with different services: LDOS CoMoDa, LJ2M and myPersonality. Two datasets were acquired in controlled, laboratory settings: LDOS PerAff-1 and DEAP. Finally, we also mention four stimuli datasets from the Media Core project: ANET, IADS, ANEW, IAPS, as well as the 1000 songs dataset. We summarise this information for a quick reference to researchers interested in using these datasets or preparing the acquisition procedure of their own.

9.1 Introduction

In order to carry out relevant research, appropriate datasets must be used, which enable researchers to test their hypotheses. For the research questions covered by this book, it is important that the datasets contain different personality-related features describing users, as well as affective parameters, along with the information regarding the interaction between users and different systems. Since the acquisition of such datasets is not an easy task, and not many live systems include such data, these datasets are rare and researchers often tend to collect their own data for research purposes.

A. Odić (✉)

Outfit7 (Slovenian Subsidiary Ekipa2 D.o.o.), Ljubljana, Slovenia

e-mail: ante.odic@outfit7.com

A. Košir

Faculty of Electrical Engineering, The User-adapted Communications & Ambient Intelligence Lab (LUCAMI), Ljubljana, Slovenia

e-mail: andrej.kosir@fe.uni-lj.si

M. Tkalčič

Department of Computational Perception, Johannes Kepler University in Linz, Linz, Austria

e-mail: marko.tkalci@jku.at

© Springer International Publishing Switzerland 2016

M. Tkalčič et al. (eds.), *Emotions and Personality in Personalized Services*, Human-Computer Interaction Series, DOI 10.1007/978-3-319-31413-6_9

In this chapter we try to provide relevant information for the researchers in the field from two perspectives: (i) survey existing and available datasets for research and (ii) survey research describing the acquisition of such datasets as a reference for the acquisition tasks and guidance for building new datasets. The datasets that we describe in this chapter are publicly available. In addition, we believe that these datasets contain valuable data for many different research goals, and as such, serve as a valuable resource for researchers. Since the acquisition of such data requires careful and controlled procedures, this section also tries to be a reference to researchers that will perform the acquisition, and preprocessing, of new datasets.

Datasets for the research in this field should ideally have several main types of information. Mainly, there has to be information regarding users, items and some type of metric that describes the interaction between users and items, how items are suitable for users or how users perceived or rated (explicitly or implicitly) the items they have consumed. This can be observed as a user-item matrix where for user-item pairs there is a measure describing their interaction (e.g. rating, different user-experience measures, etc.). In addition to that, these datasets should contain information describing users' personality profiles (e.g. Big5 factors describing users [1]). Furthermore,

Table 9.1 This table contains the list of datasets described in this section with the following information: name of the dataset, domain (i.e. type of items), whether the acquisition was in the laboratory setting or the natural interaction with the system, number of users, number of items and reference to the article describing the dataset and the acquisition

Dataset	Domain	Acquisition type	Users	Items	References
LDOS-CoMoDa	Movies	Natural interaction	235	1300	[2]
LDOS-PerAff-1	Images	Laboratory setting	52	70	[3]
LJ2M	Blogs	Natural interaction	649,712	1,928,868	[4]
DEAP	Music Videos	Laboratory setting	32	120	[5]
myPersonality	Social Network	Natural interaction	Varies	Varies	[6]
1000 songs	Music	Stimuli dataset	100	744	[7]
ANET	Text	Stimuli dataset	n/a	n/a	[8]
IADS	Sounds	Stimuli dataset	n/a	n/a	[9]
ANEW	Text	Stimuli dataset	n/a	n/a	[10]
IAPS	Images	Stimuli dataset	n/a	n/a	[11]

Table 9.2 In this table, in the same order as in the previous one, we add additional information regarding each dataset: type of personality profile of users, type of affective data, metric describing the interaction between the users and items, additional data

Dataset	Personality	Emotions	Interaction	Additional data
LDOS-CoMoDa	Big5	6 basic emotions	Ratings	Context, demographics, item metadata
LDOS-PerAff-1	Big5	VAD space	Ratings	Demographics, frontal face videos
LJ2M	No	132 mood tags	No	Blog post, associated songs
DEAP	No	VAD space	VAD ratings	Frontal face videos, sensor signals
myPersonality	Big5	No	Facebook likes	Psychometric features, demographics, etc.
1000 songs	No	VA space	VAD ratings	40-s music clips
ANET	No	VAD space	No	English texts
IADS	No	VAD space	No	Digital sounds
ANEW	No	VAD space	No	English words
IAPS	No	VAD space	No	Images

these datasets should also contain affective parameters. For example, those parameters might describe emotional states of the user during the interaction with the item, the change of the user’s emotional state after consuming an item, affective metadata describing the items, etc.

There are three types of datasets that we cover in this chapter:

1. Datasets acquired by users’ natural interaction with live systems.
2. Datasets acquired in controlled, laboratory settings, from participants.
3. Datasets containing stimuli for research on emotions.

In Tables 9.1 and 9.2 we summarize the information regarding the described datasets for quick reference.

9.2 Available Datasets

In this section we describe datasets that were made publicly available by their authors. For each dataset we provide the basic information that will help researchers to decide whether the dataset is suitable for their work. This consists of the research the dataset was intended for, description of the data the dataset contains, brief description of the acquisition procedure, and links to where the datasets can be obtained.

There are two types of datasets that are publicly available: (i) those acquired in the experimental (laboratory) setting, and (ii) those that were acquired during a natural users' interaction with an existing service. The data in the former type of datasets is usually less noisy, since all external, uncontrolled influences were removed, however such datasets contain less users due to natural limitations of the laboratory-based acquisition procedures. In addition, there is a potential problem with acquiring emotional state in laboratory setting. Emotional state of the user might be context dependent and the context in the laboratory is artificial and might not represent the real world behaviour of users. This has to be considered and addressed during the data acquisition.

On the other hand, datasets acquired in the laboratory settings have more personality and emotion-related features describing the users, since the usage of video cameras and/or other sensors were possible during the users' interaction with the system. Therefore, the selection of the dataset relevant for a specific research depends on the types of the analyses that will be performed.

While acquiring the data, it is also important to keep in mind what is user's goal and what is the value exchange for the user, especially in the case of the laboratory settings. The user's goal is the natural, or the artificial, goal that the user is trying to achieve through the interaction with the system. For example, in recommender systems, users are providing data and rating items to improve their profile in order to get more relevant recommendations. In laboratory settings, users' goals should also be explained to subjects since it is relevant whether users are rating videos, e.g. according to how suitable it is to watch at home with friends, or how interesting it is during the laboratory session. All users should be artificially placed in the same context, i.e. purpose for providing data. Regarding the value exchange, users can be motivated to use the system or participate in the experiment by different internal or external motivators, that should also be taken into account.

We mention users' goals in the description of datasets for which we found the reliable information regarding this aspect of the acquisition.

9.2.1 Context Movie Dataset (LDOS-CoMoDa)

Context Movie Dataset (LDOS-CoMoDa) was created for the research on context-aware recommender systems [2]. It was acquired from the users' natural interaction with the live system over a long period of time. It contains movie ratings, contextual information, movies' metadata and users' Big5 personality profiles from subset of users that decided to provide personality profiles.

For the data acquisition, the authors created an online application for rating movies (www.ldos.si/recommender.html). The application is used by users to track the movies they watched, obtain the movie recommendations and browse the movies.

The users were acquired by presenting and advertising the online application to students of the Faculty of Electrical Engineering, University of Ljubljana, and different movie forums and usenet newsgroups. Therefore, the users were volunteers

that were either attracted by the research questions or the usability of the online application. According to the authors, the users' goal for rating the movies is to improve their profile to gain better recommendations, express themselves and help others, according to [12].

9.2.1.1 Acquisition

Regarding the data acquisition, the online application is used by users to rate the movie that they have just seen. The users rate the items on the Likert scale from one to five.

In addition to rating the consumed movie, users fill in a simple questionnaire created to explicitly acquire the contextual information describing the situation during the consumption stage of the user-item interaction [2]. The questionnaire was designed in such a way that it is simple and not time consuming for a user to provide the contextual information. Users are instructed to provide the rating and contextual information immediately after the consumption.

Among different types of contextual information (described in the following section), emotional context was also acquired. According to the authors, for the emotional state as the contextual information in the movie RS, the consumption stage is a *multiple-context value* stage, which means that emotional state changes several times during the consumption. Therefore two types of emotional state contextual factors were acquired: (i) the emotional state that was dominant during the consumption (*domEMO*) and (ii) the emotional state at the end of the movie (*endEMO*).

Users were also able to input their personality profile into the online application. Therefore, for users that chose to do so, the dataset also contains Big5 personality profiles, that were acquired through the IPIP 50 questionnaire [13]. Ratings for movies, and all additional information was provided by users, as they have decided. There was no mandatory ratings of the preselected movies.

9.2.1.2 Dataset Information

The LDOS-CoMoDa dataset has been in development since Sep. 15, 2010. It contains three main groups of information: general user information, item metadata and contextual information. The general user information is provided by the user upon registering in the system. It consists of the user's age, sex, country and city. There are 163 male and 72 female users in the dataset.

The item metadata is inserted into the dataset for each movie rated by at least one user. The metadata describing each item is the director's name and surname, country, language, year, three genres, three actors and budget.

Table 9.3 contains the description of the acquired contextual information.

To ensure that all the acquired contextual information is from the consumption stage, the users were instructed to provide the rating immediately after the consumption, and that it should describe the moment of watching the movie. Furthermore,

Table 9.3 Contextual variables in the LDOS-CoMoDa dataset

Contextual variable	Description
time	Morning, afternoon, evening, night
daytype	Working day, weekend, holiday
season	Spring, summer, autumn, winter
location	Home, public place, friend's house
weather	Sunny/clear, rainy, stormy, snowy, cloudy
social	Alone, partner, friends, colleagues, parents, public, family
endEmo	Sad, happy, scared, surprised, angry, disgusted, neutral
dominantEmo	Sad, happy, scared, surprised, angry, disgusted, neutral
mood	Positive, neutral, negative
physical	Healthy, ill
decision	User's choice, given by other
interaction	First, n-th

this ensures that the provided rating is not influenced by unwanted noise, such as discussing the movie with friends, reading reviews, seeing the average movie rating, etc. For assessing whether the rating was provided in a satisfactory manner, the authors have identified a set of criteria that they use to flag suspicious data inputs. For example, if the rating with *winter* context is provided during summer, the data is flagged as suspicious, furthermore, if a single user provides multiple ratings at once the data is flagged as suspicious, etc. Such suspicious entries were later avoided during the testing. It is still, however, impossible to be completely sure that all the acquired data is correct. Acquiring ratings immediately after the consumption provides less noisy, real data, however, due to the collection of the contextual data, it was not possible to provide users with a list of items to rate. Each rating is made after the real consumption, which makes this type of data acquisition a long process.

LDOS-CoMoDa dataset was used in several research, for example, for the research on the role of emotions in context-aware recommendations [14], and the research on local context modelling with semantic pre-filtering [15].

Additional information about the dataset can also be found in [16]. LDOS-CoMoDa dataset can be acquired on the following link: (www.ldos.si/comoda.html).

9.2.2 LDOS-PerAff-1

The LDOS-PerAff-1 dataset was created for the need of the research of affective- and personality-based user modelling in recommender systems [3]. It is acquired from the users in the controlled, laboratory setting. LDOS-PerAff-1 dataset is composed of users' ratings for images, information about users and images, users' personality profiles, information about the induced emotions and the video clips of the users' facial expressions.

For the data acquisition, the authors created a Matlab application in which the users are rating images. The goal of rating the images was users' selection of images for their computer's wallpaper.

Users in the dataset are students that participated in the experiment. The users' goal was to rate images for the purpose of selecting the best image for their desktop background.

9.2.2.1 Acquisition

The acquisition scenario consisted of showing the subjects a sequence of images and asking the subjects to rate these images as if they were choosing images for their computer wallpaper. Ratings were selected on a Likert scale from one to five.

For each image the authors needed to know the emotional state it induces. The affective values for each image were provided by the IAPS dataset. Each image was annotated with the first two statistical moments of the induced emotion in users in the Valence-Arousal-Dominance (VAD) space [17]. The acquisition of the induced emotions was carried out by Lang et al. [18] using the Self-Assessment Manikin (SAM) questionnaire. This served as ground truth for automatic method for emotion detection and as a metadata for each image.

While users were rating images, the authors recoded their facial expressions with a camera placed on the monitor. The authors also annotated genre to each image manually through a controlled procedure.

In addition, the authors wanted to explore the relations between the subjects' personalities and their preferences for the content items. They used the IPIP 50 questionnaire to assess the factors of the Big5 factor model of the participants. The questionnaire consisted of 50 items, 10 per each of the Five-Factor Model (FFM) factors.

9.2.2.2 Dataset Information

There were 52 students who participated in the experiment. The average age was 18.3 years (standard deviation is 0.56). There were 15 males and 37 females.

The corpus consists of 3640 video clips of 52 participants responding to 70 different visual stimuli. The video files are segmented by user and by visual stimulus.

Each video clip is annotated with a line in the annotation file. The annotations are stored in text-based files. The participants cover a heterogeneous area in the space of the big five factors.

Each video clip is annotated with a line in the annotation file. The annotations files have the following columns: user id, image id, image tag, genre, watching time, wt mean, valence mean, valence stdev, arousal mean, arousal stdev, dominance mean, dominance stdev, big5 1, big5 2, big5 3, big5 4, big5 5 gender, age, explicit rating, binary rating.

For example, the LDOS-PerAff-1 dataset was used in research on using affective parameters in a content-based recommender system [19], and the research on addressing the new user problem with a personality-based user similarity measure [20].

LDOS-PerAff-1 dataset can be acquired on the following link: (<http://slavnik.fe.uni-lj.si/markot/Main/LDOS-PerAff-1>).

9.2.3 LJ2M Dataset

LiveJournal two-million post (LJ2M) dataset was collected from the social blogging service LiveJournal2 for research on personalised music-information retrieval [4]. It is acquired from the users' natural interaction with the live system over a long period of time. According to the authors, it is suitable for use in research on context-aware music recommendation, emotion-based playlist generation, affective human-computer interface and music therapy.

LJ2M dataset contains a blog article, a song associated with the post, and a user mood, since each article is accompanied with a mood and music entries which the blog authors provide.

Users in the dataset are bloggers that use LiveJournal social-networking service. The users' goal was blogging about different subjects.

9.2.3.1 Acquisition

LiveJournal is a social-networking service with a large user base (according to the authors, 40 million registered users and 1.8 million active users at the end of 2012). As described in [21], for purposes of sentiment analysis Leshed and Kaye collected 21 million posts using the LiveJournal's RSS feeds. Maximum of 25 posts were collected per user. Users are mostly from United States, and articles were written between 2000 and 2005.

Each LiveJournal's post contains an article, a mood entry and a music entry. Mood entries were provided by the blog-article's author by selecting one of the 132 pre-defined tags, or filling in freely. Similarly, blog-article authors also provided the music entry by filling in anything they wish.

The authors in [4] further processed the raw data acquired in [21]. They considered only those entries that contained pre-defined mood tags. Regarding the music entry, they used the AchoNest API to check the existence in the EchoNest database, and considered those entries with valid and found (artist, song title pairs). Finally, only those blog entries that contained both valid mood and music tag were selected. The content of the articles is provided as lists of word counts with both non-stemmed and stemmed versions.

9.2.3.2 Dataset Information

The dataset contains 1,928,868 posts from 649,712 unique users. There are $14,613 \pm 13,748$ articles per mood tag, on average. Majority of the mood tags have more than 1,000 articles, and about half have more than 10,000 articles.

Blog articles contain 88,164 unique songs from 12,201 artists. There are 125 ± 22 articles per song, on average. 64,124 songs can be found in the million song dataset (MSD) [22], hence musical metadata and features from MSD can be used. 87,708 songs have short audio previews (30s) available from 7digital.

LJ2M is available at <http://mac.citi.sinica.edu.tw/lj/>.

9.2.4 A Database for Emotion Analysis Using Physiological Signals (DEAP)

The Database for Emotion Analysis Using Physiological Signals (DEAP) is a multi-modal data set for the analysis of human affective states [5]. This dataset was acquired in a controlled laboratory setting. It contains the electroencephalogram (EEG) and peripheral physiological signals of users, their ratings of arousal, valence, like/dislike, dominance and familiarity of music videos presented, frontal face video recording for a subset of the participants, subjective ratings from the initial online subjective annotation and the list of 120 videos used.

The authors prepared a laboratory setting in which users watched 40 1-min long excerpts of music videos, provided ratings in terms of arousal, valence, like/dislike, dominance and familiarity, while different signals were taken from them through sensors.

Users in this dataset are volunteers that participated in the experiment.

9.2.4.1 Acquisition

120 music videos as emotional stimuli were selected, and 1-min segments with highest emotional content were extracted. Through a web-based assessment experiment participants rated the 1-min segments on a discrete 9-point scale for valence, arousal and dominance. After each video segment was rated by at least 14 volunteers, 40 videos were selected for use in the experiment. To achieve maximum strength of elicited emotions, selected videos had the strongest volunteer ratings and smallest variation.

Once the 1-min-video stimuli was selected the experiment was prepared. Participants were prepared and instructed, set in a controlled environment and sensors were placed on them. The experiment started with a 2-min baseline recording after which 40 videos were presented in 40 trials. In each trial, a 2-s screen displaying the current trial number was shown to inform the participant of the progress, followed by the 5-s

baseline recording, and finally the 1-min music-video stimuli. After the video was shown participants performed a self-assessment of their levels of arousal, valence, dominance via the self-assessment manikins on a continuous scale. In addition, they rated how much they liked the shown video by the thumbs-down and thumbs-up symbols.

9.2.4.2 Dataset Information

The dataset consists of two parts: (i) online subjective annotation and (ii) psychological experiment.

Subjective Annotations. There are 120 1-min music videos, 60 of which were selected via last.fm [23] affective tags and 60 were selected manually. Each video has 14–16 ratings on arousal, dominance and valence discrete scale of 1–9.

Physiological Experiment. Thirty-two participants, 50 % of which are females, aged between 19 and 37, rated 40 1-min videos. Ratings were made on scales: arousal, dominance, valence, liking and familiarity. Familiarity is rated on a discrete scale from 1 to 5, and other ratings on a continuous scale from 1 to 9.

In addition to ratings, dataset contains 32-channel 512 Hz EEG signals, and peripheral physiological signals. For 22 participants dataset contains frontal face video.

The article [5] contains detailed information about the data acquisition as well as the analysis on the acquired data. DEAP dataset is published and available for research on the following link:

For example, DEAP dataset was used in research on multi-task and multi-view learning of user state [24], and EEG-based Emotion Recognition by using deep learning network [25].

(<http://www.eecs.qmul.ac.uk/mmv/datasets/deap/index.html>).

9.2.5 *myPersonality Project Dataset*

The myPersonality project dataset is a dataset that can be used for different research tasks on social network behaviour and users in connection to different psychometric features [6]. This dataset is acquired from users' natural interaction with the Facebook application, and their natural interaction with social network Facebook. It contains data regarding Facebook users, their preferences (Facebook likes), various demographic information, as well as psychometric data from different tests users have participated in.

The data was acquired by the myPersonality Facebook application that allowed users to take real psychometric tests. If users so decided, they could also provide different Facebook data from their Facebook profiles.

Users in this dataset are therefore Facebook users that decided to use myPersonality application. The users goal was to get feedback and interesting information from different tests in the application.

9.2.5.1 Acquisition

The data was acquired by the myPersonality application (www.mypersonality.org). Users provided their data and gave consent to have their scores and profile information recorded. There were two acquisition principles in this project: (i) acquiring data from psychometric tests and surveys and (ii) collecting users’ Facebook data that users have shared.

Political and religious views, sexual orientation and relationship status were recorded from the related fields of users’ Facebook profiles. Ethnicity was added in form of labels which were assigned to users by visual inspection of their profile pictures. According to the authors, this resolves the problem of the disclosure bias, however, not all users had profile pictures showing themselves. Information regarding substance use and whether users’ parents stayed together or split up before the user was 21 years old were acquired by self-report survey on the myPersonality application. User’s Personality Five-Factor Model (FFM) was acquired by the International Personality Item Pool questionnaire with 20 items. User’s intelligence was measured by Ravens Standard Progressive Matrices, which is a multiple choice nonverbal intelligence test based on Spearman’s theory of general ability. Users’ satisfaction with life was measured using a popular, five-item SWL Scale, which measure global cognitive judgments of satisfaction with one’s life. The author also recorded more than 9 million unique objects liked by users, but have removed likes associated with fewer than 20 users, as well as users with fewer than two likes.

9.2.5.2 Dataset Information

The myPersonality project dataset contains many different variables describing users. However, not all variables are available for all users. In Table 9.4 we show the approx.

Table 9.4 Number of records for selected subset of variables in myPersonality project dataset

Variable	Approx. number of records
User’s demographic details	4,300,000
User’s geo-location details	1,800,000
Facebook activity	1,600,000
User’s religion and political views	331,672
BIG5 Personality Scores	3,100,000
IQ scores	7,000
Satisfaction with life scale	101,000
Barratt impulsivity scale	19,000
Body consciousness questionnaire	14,000
Facebook likes dictionary	5,500,000
Photos dictionary	17,200,000
Schools dictionary	128,000
Users’ Facebook status updates	22,000,000 updates of 154,000

number of records, i.e. users in some cases, for which a specific variable is available. Since there are many variables, we select some of them to give readers the general idea about the number of records. All information on all variables can be found on the following link: <http://mypersonality.org/>.

All the additional information about the myPersonality project dataset can be found on the following link: <http://mypersonality.org/>. On the same link, after registering as a collaborator, it is possible to obtain various parts of the dataset.

The myPersonality dataset was used in many research, for example, research on the automatic personality assessment through social media language [26], and relating personality types with user preferences [27].

9.3 Stimuli Datasets

In this section we present stimuli datasets with various types of items which can be used in the research of emotions.

9.3.1 *Emotion in Music Database (1000 Songs)*

Emotion in Music Database is a stimuli dataset that can be used for the development of music emotion recognition systems [7]. It contains songs and affective annotations provided by volunteers. Each song is annotated with valence and arousal, both continuously, throughout the duration of the song, and statically at the end of the song.

The authors developed the online-annotation system which volunteers were using for the task.

9.3.1.1 Acquisition

The authors first acquired 1000 Creative Commons (CC) licenced music from the Free Music Archive (FMA) [28]. 125 songs were selected from each of the eight different genres: Blues, Electronic, Rock, Classical, Folk, Jazz, Country and Pop. From the initial larger sample of songs, all those that were longer than ten and shorter than one minute were excluded.

The authors were interested in annotating each song with the valence–arousal annotations from multiple annotators. In addition, two different annotations were made, time-varying (per second) continuous valence–arousal ratings, and a single discrete (9 point) valence–arousal rating applied to the entire clip. For the task, the authors have developed their own online-annotation interface for music. Via the interface, annotators are continuously annotating each song during listening by the slider indicating the current emotion. After annotating the songs continuously,

annotators are additionally asked to rate the level of arousal or valence for the whole clip on a 9 point scale through Self-Assessment Manikins.

Annotators were acquired by a crowdsourcing principle using the Amazon Mechanical Turk. To ensure the quality of annotators, the authors designed a quality control strategy for the acquisition of annotators.

9.3.1.2 Dataset Information

Initially, the dataset contained 1000 40-s clips, and each clip was annotated by a minimum of 10 workers. More than 20,000 annotations were collected. From 100 workers who participated in the annotation procedure, 57 were male and 43 were female, with average age of 31.7 ± 10.1 . On average, annotators spent 7 min and 40 s annotating three clips. Annotators were from 10 different countries, 72 % from the USA, 18 % from India and 10 % from the rest of the world.

The authors found redundant songs and cleaned the data which reduced the number of songs down to 744.

The audio files are distributable under the CC licence and can be shared freely. FMA songs are not published by music labels so the annotators are usually less familiar with them and the potential biases introduced by familiarity with the songs are reduced.

For example of usage, 1000 songs dataset was used in research on emotional analysis of music [29], and continuous-time music mood regression [30].

1000 Songs dataset is published and can be acquired at the following link: <http://cvml.unige.ch/databases/emoMusic/>

9.3.2 Media Core

Media Core is a project of the Centre for the Study of Emotion and Attention, University of Florida, which develops, catalogues, evaluates and distributes various types of media (stimuli) that can be used as prompts to affective experience [31]. In this section, we describe and provide links to four stimuli datasets from Media Core which cover: images, sounds, English words and English texts.

The Affective Norms for English Text (ANET) dataset provides a set of emotional stimuli from text [8]. It contains a large set of brief English texts. Each text is accompanied by the normative rating of emotion in terms of valence, arousal and dominance dimensions. ANET dataset is publicly available and accessible on the following link: <http://csea.php.ufl.edu/media/anetmessage.html>.

The International Affective Digital Sounds (IADS) dataset provides a set of emotional stimuli from digital sound [9]. It contains a large set of digital sounds accompanied by the normative rating of emotion in terms of valence, arousal and dominance dimensions. IADS dataset is publicly available and accessible on the following link: <http://csea.php.ufl.edu/media/iadsmessage.html>.

The Affective Norms for English Words (ANEW) dataset provides a set of emotional ratings for a large set of English words [10]. Each word is accompanied by the normative rating of emotion in terms of valence, arousal and dominance dimensions. ANEW dataset is publicly available and accessible on the following link: <http://csea.phhp.ufl.edu/media/anewmessage.html>.

The International Affective Picture System (IAPS) provides a set of emotional stimuli from images [11]. It contains a large set of colour photographs which cover a wide range of semantic categories. Each image is accompanied by the normative rating of emotion in terms of valence, arousal and dominance dimensions. IAPS dataset is publicly available and accessible on the following link: <http://csea.phhp.ufl.edu/media/iapsmessage.html>.

9.4 Conclusion and Summary

In this chapter we presented some of the available datasets which contain information relevant to the research questions addressed in this book. All datasets are publicly available for researchers working in the field.

For each dataset we briefly described several aspects that we find important for the decision whether to use the dataset, as well as for the reference for researchers interested in acquiring new datasets. Therefore, we describe the acquisition procedure, data the datasets contain, links to the datasets, examples of research done on these datasets, where applicable description of users' goal during the acquisition and additional specific information.

In our opinion, researchers that are planning to acquire new datasets, relevant to this field of research, should pay special attention to carefully specifying users' goals and the context in which users are providing the data. In addition, reduction of noise in the data should be considered and employed for which ideas and procedures can be found in the articles associated to the datasets described in this chapter.

Unfortunately, there is still a low number of publicly available datasets relevant to the field. This is due to complex procedures needed for the data acquisition in laboratory settings and potentially sensitive personal information in real-live systems. However, with increasing accessibility of available sensors and stimuli datasets as well as crowdsourcing platforms and overall usage of affective and personality data used in existing services, amount of data relevant for these research topics is also on the rise. We hope that the information and references from this chapter will help researchers to find the appropriate datasets for their work, provide useful information for the preparation their own acquisition procedures, and motivate them to share their datasets with other researchers in this field of research.

References

1. Saucier, G., Goldberg, L.R.: What is beyond the big five? *J. Pers.* **66**, 495–524 (1998)
2. Odić, A., Tkalčić, M., Tasić, J.F., Košir, A.: Predicting and detecting the relevant contextual information in a movie-recommender system. *Interact. Comput.* **25**(1), 74–90 (2013)
3. Tkalčić, M., Košir, A., Tasić, J.: The LDOS-PerAff-1 corpus of facial-expression video clips with affective, personality and user-interaction metadata. *J. Multimodal User Interface* **7**(1–2), 143–155 (2013)
4. Liu, J.Y., Liu, S.Y., Yang, Y.H.: LJ2M dataset: toward better understanding of music listening behavior and user mood. In: *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6 (2014)
5. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Patras, I.: Deap: a database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **3**(1), 18–31 (2012)
6. Kosinski, M., Stillwell D.J., Graepel T: Private traits and attributes are predictable from digital records of human behavior. In: *Proceedings of the National Academy of Sciences (PNAS)* (2013)
7. Soleymani, M., Caro, M.N., Schmidt, E.M., Sha, C.Y., Yang, Y.H.: 1000 songs for emotional analysis of music. In: *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, pp. 1–6 (2013)
8. Bradley, M.M., Lang, P.J.: *Affective Norms for English Text (ANET): affective ratings of text and instruction manual*. (Tech. Rep. No. D-1). University of Florida, Gainesville, FL (2007)
9. Bradley, M.M., Lang, P.J.: *International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings* (Tech. Rep. No. B-2). The Center for Research in Psychophysiology, University of Florida, Gainesville, FL (1999)
10. Bradley, M.M., Lang, P.J.: *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*. Technical report C-1. The Center for Research in Psychophysiology, University of Florida, Gainesville, FL (1999)
11. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: *International affective picture system (IAPS): affective ratings of pictures and instruction manual*. Technical Report A-8. University of Florida, Gainesville, FL (2008)
12. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. (TOIS)* **22**(1), 5–53 (2004)
13. Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G.: The international personality item pool and the future of public-domain personality measures. *J. Res. Pers.* 84–96 (2006)
14. Zheng, Y., Mobasher, B., Burke, R.D.: The role of emotions in context-aware Recommendation. In: *Decisions@ RecSys*, pp 21–28 (2013)
15. Codina, V., Ricci, F., Ceccaroni, L.: Local context modeling with semantic pre-filtering. In: *Proceedings of the 7th ACM Conference on Recommender Systems*, pp. 363–366 (2013)
16. Košir, A., Odić, A., Kunaver, M., Tkalčić, M., Tasić, J.F.: Database for contextual personalization. *Elektrotehniki vestnik* **78**(5), 270–274 (2011)
17. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **17**(03), 715–734 (2005)
18. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. Technical Report A-8 (2008)
19. Tkali, M., Burnik, U., Koir, A.: Using affective parameters in a content-based recommender system for images. *User Model. User-Adap. Inter.* **20**(4), 279–311 (2010)
20. Tkalčić, M., Kunaver, M., Koir, A., Tasić, J.: Addressing the new user problem with a personality based user similarity measure. In *First International Workshop on Decision Making and Recommendation Acceptance Issues in Recommender Systems (DEMRA 2011)* (2011)
21. Leshed, G., Kaye, J.J.: Understanding how bloggers feel: recognizing affect in blog posts. In: *CHI'06 extended abstracts on Human factors in computing systems*, pp. 1019–1024 (2006)

22. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The million song dataset. In: ISMIR: Proceedings of the 12th International Society for Music Information Retrieval Conference, Miami, Florida, vol. 591–596 (2011)
23. <http://www.last.fm/>
24. Kandemir, M., Vetek, A., Gnen, M., Klami, A., Kaski, S.: Multi-task and multi-view learning of user state. *Neurocomputing* **139**, 97–106 (2014)
25. Jirayucharoensak, S., Pan-Ngum, S., Israsena, P.: EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *Sci. World J.* (2014)
26. Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Seligman, M.E.: Automatic Personality Assessment Through Social Media Language (2014)
27. Cantador, I., Fernández-Tobas, I., Bellogn, A., Kosinski, M., Stillwell, D.: Relating personality types with user preferences in multiple entertainment domains. In: UMAP Workshops (2013)
28. <http://www.freemusicarchive.org><http://www.freemusicarchive.org>
29. Soleymani, M., Aljanaki, A., Yang, Y.H., Caro, M.N., Eyben, F., Markov, K., Wiering, F.: Emotional analysis of music: a comparison of methods. In: Proceedings of the ACM International Conference on Multimedia, pp. 1161–1164 (2014)
30. Weninger, F., Eyben, F., Schuller, B.: On-line continuous-time music mood regression with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5412–5416 (2014)
31. <http://csea.php.ufl.edu/media.html>