석사학위논문
Master's Thesis

능동 학습을 통한 실시간 감정의 적응적 샘플링

Towards Adaptive Sampling of Emotions in the Wild with Active Learning

2021

박 철 영 (朴 哲 永 Park, Cheul Young)

한국과학기술원

Korea Advanced Institute of Science and Technology

석사학위논문

능동 학습을 통한 실시간 감정의 적응적 샘플링

2021

박철영

한국과학기술원

지식서비스공학대학원

능동 학습을 통한 실시간 감정의 적응적 샘플링

박 철 영

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2020년 12월 23일

심사위원장	이 의 진	(인)
심 사 위 원	이 문 용	(인)
심 사 위 원	Ahsan Habib Khandoker	(인)

Towards Adaptive Sampling of Emotions in the Wild with Active Learning

Cheul Young Park

Advisor: Uichin Lee

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Knowledge Service Engineering

Daejeon, Korea
December 23, 2020

Approved by

Uichin Lee
Professor of Computer Science

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MKSE
20184316

박철영. 능동 학습을 통한 실시간 감정의 적응적 샘플링. 지식서비스공학대학원 . 2021년. 42+iv 쪽. 지도교수: 이의진. (영문 논문)
Cheul Young Park. Towards Adaptive Sampling of Emotions in the Wild with Active Learning. Graduate School of Knowledge Service Engineering . 2021. 42+iv pages. Advisor: Uichin Lee. (Text in English)

초 록

일상적으로 발생하는 감정의 인식 기술은 다양한 활용처가 있지만 정확한 인식 기술의 개발은 자연 발생하는 감정들의 수집에 필요한 높은 비용과 감정들의 불균형한 분포로 인해 제한된다. 지금까지 사용되어온 임의로 감정을 수집하는 방식은 위 문제들의 해결에 적합하지 않다. 따라서 이 논문에서는 능동 학습을 통한 감정의 적응적 수집 방법을 수집되는 감정의 분포를 더 고르게 하고 수집 대상자의 부담을 줄일 수 있는 대안으로 제안한다. 또한, 적응적 감정 수집 방식의 효과를 자연적인 대화 상황을 가정한 환경에서 생체 신호와 연속적인 감정을 수집한 K-EmoCon 데이터셋을 사용한 실험으로 살펴보며, 능동 학습에 있어 균형 잡힌 데이터의 수집과 학습에 유익한 샘플의 수집 간의 상관관계를 매개변수를 활용한 질의 전략으로 살펴본다.

핵 심 낱 말 인간-컴퓨터 상호작용, 감성 컴퓨팅, 경험 샘플링 방법, 기계 학습, 능동 학습

Abstract

Accurate recognition of emotions has many applications, but it is challenged by difficulties in collecting emotions in the wild. While naturally occurring emotions are expensive to collect, the inherent bias in their distribution further confounds the issue. The random sampling method frequently employed by researchers fails to overcome these limitations. We propose an adaptive sampling method based on active learning as an alternative to collect emotions with balanced distribution while reducing burdens on users. The effectiveness of adaptive sampling is empirically evaluated with the K-EmoCon, the dataset of continuous emotions and physiological signals collected in the context of naturalistic conversations. The tradeoff between collecting balanced data and querying informative samples is also explored with a parameterized query strategy.

Keywords human-computer interaction, affective computing, experience sampling method, machine learning, active learning

Contents

Contents . . . i

List of Tables . . . iii

List of Figures . . . iv

Chapter 1. Background & Summary **1**

Chapter 2. Related Works **4**

2.1 Theories of Emotion and Recent Advances in the Field . . . 4

Chapter 3. Data and Methods **6**

3.1 Data Description . . . 6

3.1.1 Data collection . . . 7

3.1.2 Dataset contents . . . 10

3.2 Data Preprocessing . . . 10

3.2.1 Feature extraction from physiological signals . . . 10

3.2.2 Deep networks for physiology-based emotion recognition 12

3.3 Experiment Setup . . . 13

3.3.1 Binary classification of emotions . . . 14

3.3.2 Query strategy for active learning . . . 14

3.3.3 Online active learning . . . 17

3.3.4 Evaluation metrics for imbalanced data . . . 19

Chapter 4. Experiments and Results **20**

4.1 Baseline Classification . . . 20

4.1.1 XGBoost . . . 20

4.1.2 LSTM . . . 21

4.2 Simulated Stream-based Active Learning . . . 21

4.2.1 Query selectivity and model performance . . . 23

4.2.2 Tradeoff between uncertainty and minority sampling . . 24

4.2.3 Rebuilding models per updates . . . 26

4.2.4 Deep active learning . . . 28

Chapter 5. Discussion: Limitations and Future Works **29**

Chapter 6. Conclusion **31**

Acknowledgments 42

List of Tables

3.1 Summary of multimodal emotion datasets 8

3.2 Wearable sensors for collecting physiological signals in K-EmoCon 9

3.3 Contents of the K-EmoCon dataset 10

3.4 Features extracted from physiological signals 11

4.1 Baseline classification results with XGBoost 20

4.2 Baseline classification results with LSTM 22

List of Figures

2.1 Russell's circumplex model of affect 5

3.1 Data collection session setup 6
3.2 Screen capture of a debate footage 7
3.3 Wearable sensors used for the K-EmoCon construction 9
3.4 Resampling techniques for BVP and ECG 13
3.5 The distribution of self-reported arousal and valence labels in the K-EmoCon dataset . . . 13
3.6 Query probability functions for uncertainty sampling. 15
3.7 Query probability functions for paremeterized uncertainty-minority sampling. 16

4.1 Confusion matrices for the baseline classification with XGBoost. 21
4.2 Confusion matrices and train/validation losses for the baseline classification with LSTM. . 22
4.3 Learning trajectories of XGBoost-based active learners without target coverage and 60s input features. 23
4.4 Learning trajectories of XGBoost-based active learners with target coverage = 60%. . . . 24
4.5 Learning trajectories of XGBoost-based active learners at varying levels of α and β, with $\gamma = 1$ and $\phi^* = 0.6$. 25
4.6 Learning trajectories of XGBoost-based active learners at varying levels of α and β, with $\gamma = 0.1$ and no target coverage. 25
4.7 Learning trajectories of XGBoost-based active learners with model rebuilds after each update. 27
4.8 Train and validation losses of LSTM-based active learners for arousal classification. 28

Chapter 1. Background & Summary

The goal of automatic emotion recognition is to endow machines with emotional intelligence [88, 124], to allow them to understand human emotions. Such ability of machines can have many uses, ranging from medical applications including the prognosis, diagnosis, and treatment of mental illnesses [87, 108, 138], the development of more human-like virtual intelligent assistants capable of recognizing, understanding, and even possibly expressing emotions [101, 115], to the addition of simple emotional skills to smart home appliances to lubricate their social interaction with users [14, 75, 116]. This idea of giving a machine a quality that is most human of all has been a long-sought goal throughout our scientific endeavor and recently developed into a branch of its own in computer science since Picard established the field of Affective Computing [107]. So far, many researchers have contributed to the field to develop systems capable of recognizing emotions from multiple modalities, which include facial expressions [89, 94, 96], utterances [113, 140], gestures [51, 72], writings [4, 146], and physiological signals [18, 99, 110], and recently with the rapid development in the field of Artificial Intelligence, in the particular lead by advances in deep neural networks, the field is flourishing with algorithms capable of accurate detection of human affective states.

Nonetheless, the problem of emotion recognition is far from being solved, given the inherent perplexity of emotions and limitations in the real-world application of the emotion recognition technology. While the detection of explicitly observable manifestations of emotions such as facial expressions and speech is straightforward [16] and often systems based on those modalities are sufficient to be commercialized as software such as MS Azure facial recognition API and Affective Media Analytics tool, their usage is limited to narrow target scenarios, as certain configurations of facial muscles or timbre of speech do not capture the entirety of a person's internal state. In the wild, human facial expressions and speeches are intentionally regulated often to deter the decoding of underlying the affective state [46], which is entirely natural from an evolutionary and social standpoint [11, 73]. Our emotions are not absolute nor discrete nor 2-dimensional; emotions are more than a number of expressions on our faces [7]; possibly emotions are not even countable, without an injective mapping to individually discernible biophysical states, and are only definable in the context of social interactions and relationships which the emotions occurred.

On top of these complications associated with emotions, naturalistic emotion data is notoriously difficult to acquire compared to other types of data such as images of cars or human writing in different languages. At the root of this problem is the ambiguity in what constitutes as ground truth for emotions [28]. Broadly, there are two ways for annotating emotions: with self-reported labels from subjects who experience emotions first-hand or with perceived labels from external observers [154]. If these two labels concur, then we have a ground truth, but the agreement is rare for emotions naturally occurring in interactions involving multiple individuals. Another issue is the difficulty of capturing natural emotions in the wild. To capture emotions, researchers need to interrupt people when they experience or observe emotions to produce annotations or ask to assess emotions retrospectively. Nevertheless, both in-situ and retrospective approaches to emotion annotation are subject to biases [78, 79, 120]. When asked to report their emotions at the moment, people may be urged to censor negative emotions to appear more socially favorable [20]. This social desirability bias similarly influences emotions reported in retrospect, which are also subject to human memory's fallibility. Of course, information associated with emotion labels such as facial expressions need to be gathered together to develop emotion recognition systems, and this cannot avoid the issue of privacy violation.

Given these obstacles, collecting emotions in controlled environments such as laboratories with stimuli designed to induce specific emotions has been the most widely employed approach for emotion research [1, 30, 54, 65, 74, 85, 90, 131, 134, 136], as this approach can mitigate the issue of ground truth and cognitive biases in emotion labels. Simultaneously, the concern for privacy violation can be reduced with the explicit approval of subjects for data collection. However, again, the generalizability of emotion recognition systems trained with data collected in controlled environments is questionable if such systems are to be used in realistic situations. As noted, emotions are highly context-dependent [8, 21, 22] and also subject to individual variability [53, 76]. The modality of inputs to emotion recognition systems is also an issue, as it is simply implausible to monitor someone's facial expression or speeches with current technologies without violating the privacy of others.

Altogether, the issues in collecting naturalistic emotions for developing a system that can recognize emotions in a real-time call for an approach capable of probing emotions at opportune moments without imposing too much burden on subjects and involves a noninvasive method for collecting affective information associated with emotions. In light of that, Experience Sampling Methods or Ecological Momentary Assessments (ESM/EMA) [33, 77, 132] is widely used in recent years for emotion-related research [52, 97, 143], not only as a tool for observation and data collection but as means of intervention as well [59, 102, 147], together with passive data collection from mobile and wearable devices [3, 84, 95, 109, 145] or even social networks [121]. While the reliability of self-reported emotions via ESM remains an issue, ESM is well received in the field as it supports the assessment of emotions in the circumstances inaccessible with traditional methods and studying emotions in dynamically changing contexts involving varying interactions and events, thus expanding the domain of emotion research beyond the laboratories to the real world.

In the real-world adoption of the ESM for emotion research, however, major pitfalls of the method are the exhaustive data collection process compromising the ecological validity of research and the measurement effect arising from repeated answering of ESM questionnaires affecting emotions [23, 35, 126]. Further, the effect of an ESM study on what it intends to measure, i.e., emotions or symptoms of an ailment, and the quality of collected data are subject to numerous variables, including the design of an experiment and a survey instrument [56], the sampling frequency and the survey length [38], the total duration of the experiment [141], and in particular, the timeliness of an ESM survey, not only because when the measurements are taken relates to what emotions are collected [37], but the distribution of emotions in the wild is inherently imbalanced [153]. While previous studies employed various approaches to collect emotions in the wild, from the frequently employed randomly triggered [61, 91] or event-triggered questionnaires [60, 71, 81, 137], all methods are subject to biases from the underlying distribution of emotions, self-selection, and attrition.

An adaptive experience sampling method that does not rely on a heuristic to trigger questionnaires in that regard can be an alternative to traditional approaches for the in-situ collection of emotions. In part motivated by the interruptibility research [45, 111] that aims to allow computer systems to discover opportunities to interrupt human users at moments that are most appropriate for an interaction, instead of manually scheduling ESM questionnaires, an adaptive ESM would use an algorithm with specific objectives to determine the most opportune moments for sampling emotions. This idea has been explored with a rule-based approach balancing interference and data fidelity [50], annotation prediction based on active learning and semi-supervised learning [83], modeling of interruptibility from mobile phone usage and contextual information [106], and decision-theoretic models built upon predictive models [64, 117].

In the same vein, this work investigates the possibility of applying an adaptive experience sampling method for collecting emotions in the wild. Towards that goal, we use the K-EmoCon [105], a multimodal emotion dataset of continuous emotions in during debates, and train physiology-based emotion recognition models with physiological signals and continuous emotions annotations in the dataset via active learning [128] in an environment simulating a realistic emotion data collection scenario. In that process, we explore the prospects of active learning for overcoming the challenges proposed by the imbalanced distribution of naturalistic emotions in training emotion recognition models and evaluate the effectiveness of active learning for reducing the burden of users in emotion data collection. Further, we observe how different configurations of query strategies in active learning affect the performance of a resulting emotion recognition model and their implications in learning from imbalanced data, using a parameterized query strategy combining uncertainty sampling and minority sampling with additional parameters γ and ϕ controlling for the selectiveness and the coverage of an active learner. In summary, the contributions of this work are as follows:

- We extend the research of experience sampling for emotions in the wild by assessing the viability of active learning for adaptive sampling of emotions in the realistic emotion data collection scenario and demonstrate that a comparable performance can be obtained for emotion recognition models by using only 60% of the full dataset with active learning, which translates to *the possibility of a significantly reduced number of queries and user burdens in emotion data collection.*

- We assess the technical validity of K-EmoCon, a recently published dataset of multimodal affect information and continuous emotion annotations in emotion recognition research, by developing *physiology-based emotion recognition models with potential applications in real-world scenarios.*

- We explore the implications of different query strategies, especially the query selectivity of a model and the tradeoff between uncertainty and minority sampling, by training active learning models for emotion recognition using the naturalistic emotion dataset with imbalanced distributions and the parameterized query strategy, and discusses the *importance of careful exploration during the early stage of training an active learner.*

Chapter 2. Related Works

2.1 Theories of Emotion and Recent Advances in the Field

Despite the common occurrence of emotions in daily lives, there is no consensus on the scientific definition of emotions. They are often confused with mood, temperament, personality, and other psychological constructs, while researchers differentiate between emotions and moods/sentiments/traits, as emotions are intentional, i.e., object-directed, and moods are not, for example, depression [13, 48]. Moods also tend to last longer and affect our cognitive strategies, while emotions bias our actions [34]. Sentiments, on the other hand, are assigned properties of an object rather than states of an individual, like how we would say a movie was "great" or "disappointing." Then Basic Emotion Theory (BET) of Ekman [39, 40] argues for 6 basic and universal emotions of happiness, surprise, fear, sadness, anger, and disgust [42]. The term "basic" indicates that the six emotions are separate in terms of physiology, appraisal, antecedents, response, and exist across species, and "universal" as they exist across cultures.

Other theories model emotions using a number of dimensions, from Russell's Circumplex Model of Affect with two dimensions of arousal and valence [119], a similarly two-dimensional Plutchik's Wheel of Emotions [112], to the Hourglass Model of Emotions with four dimensions of pleasantness, sensitivity, aptitude, and attention [19]. The most widely used among the dimensional models of emotions is the 2D circumplex of affect, given the succinctness of the 2-D space and ease of use for simple surveys.

Nonetheless, the latest findings in emotion research suggest that emotions are not restricted to six discrete categories nor several quantifiable dimensions. Recent research on facial expressions shows that emotions are compound [36] — consider how one might simultaneously feel sad and happy on some occasions, which we informally refer to as a "sappy" feeling. Another research suggests that emotions are ordinal, i.e., relative rather than absolute [149], despite absolute annotations of emotions prevailing in the literature.

More recent research in emotion research then suggests over 20 categories of emotions exist beyond the basic six [29, 32, 67, 69, 133], including Ekman's basic six and more abstract emotions such as pride, awe, and love, and proposes that we build a high-dimensional taxonomy of emotional experiences and expressions [31, 66] using data-driven methods to capture myriads of emotional behaviors, experiences, and expressions [63]. Yet, it must be noted that these advances do not reject previous theories, including BET, as they provide a generative framework rather than a definitive framework for emotion research [70], claiming that basic emotions are not single affective states but a "family of related states." [41] Further, the discussion is ongoing, with the advent of social functionalism [68, 73, 142], which views emotions to serve distinct purposes such as survival and reproduction in society and as means of communicating information, to reward behaviors, and elicit certain behaviors during interpersonal interactions.

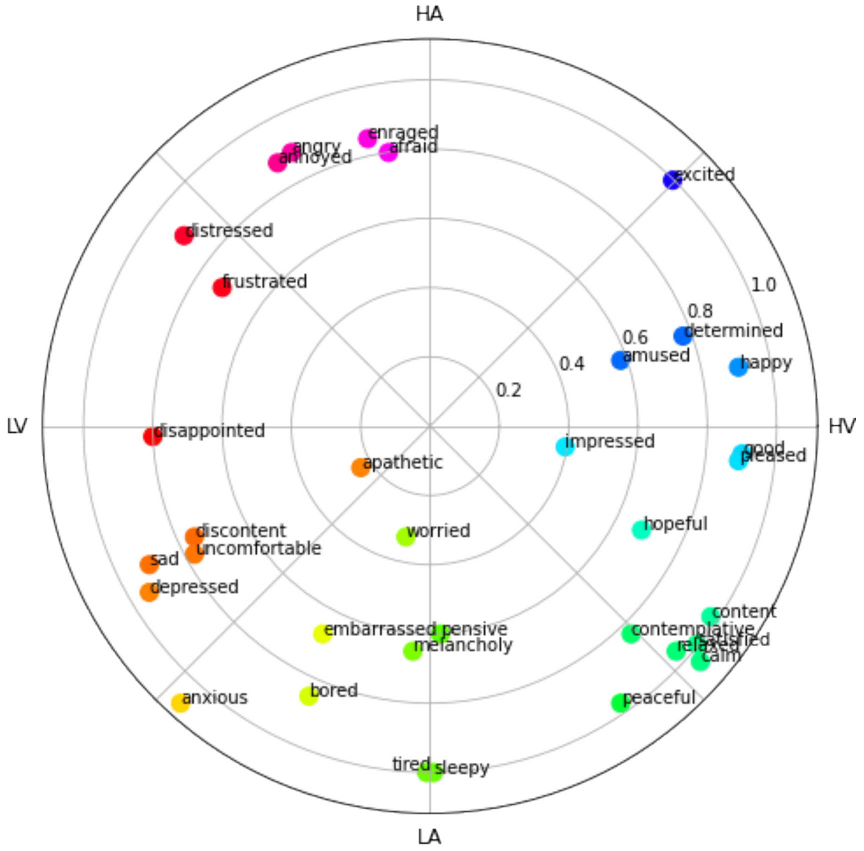


Figure 2.1: A diagram of an arousal-valence circumplex with approximate placements of emotional words on the space [103, 125].

Chapter 3. Data and Methods

In this chapter, we first discuss the data we use to develop and evaluate an adaptive sampling method for emotions and the techniques for preprocessing physiological sensor data. We then formalize emotion recognition as a classification problem in the context of adaptive sampling and propose online active learning strategies based on a parametric query function for sampling/learning imbalanced emotions in the wild. Finally, we discuss how we empirically compare active learning models' performance against offline models trained with full data and our experiments and evaluation strategies to investigate the tradeoff between collecting balanced data and querying informative samples in active learning.

3.1 Data Description

For the implementation and evaluation of an adaptive sampling method for emotions in the wild, we use the K-EmoCon dataset of physiological signals and continuous emotion annotations collected in the context of naturalistic conversations. In the following, we discuss the data collection process and its contents in detail.

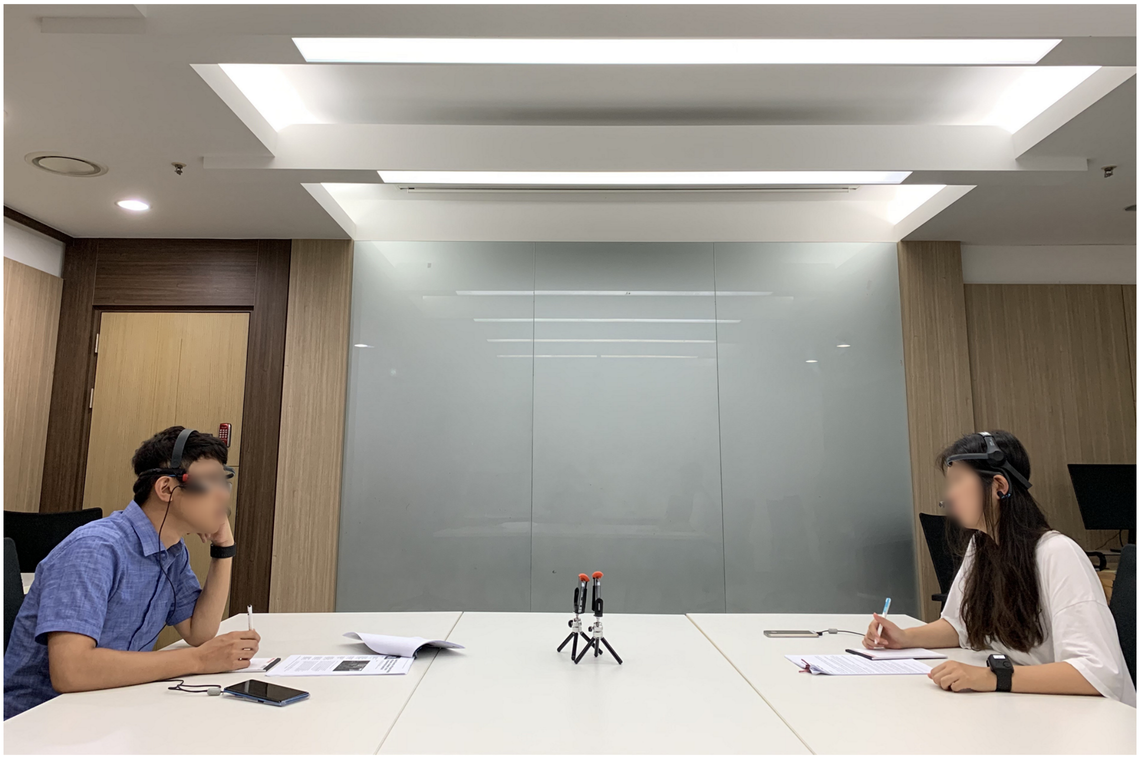


Figure 3.1: A pair of participants sitting at a table during a debate: two smartphones on tripods were placed at the center of the table to capture facial expressions and gestures in the upper body during a debate. Participants' faces are occluded per their request for privacy.

3.1.1 Data collection

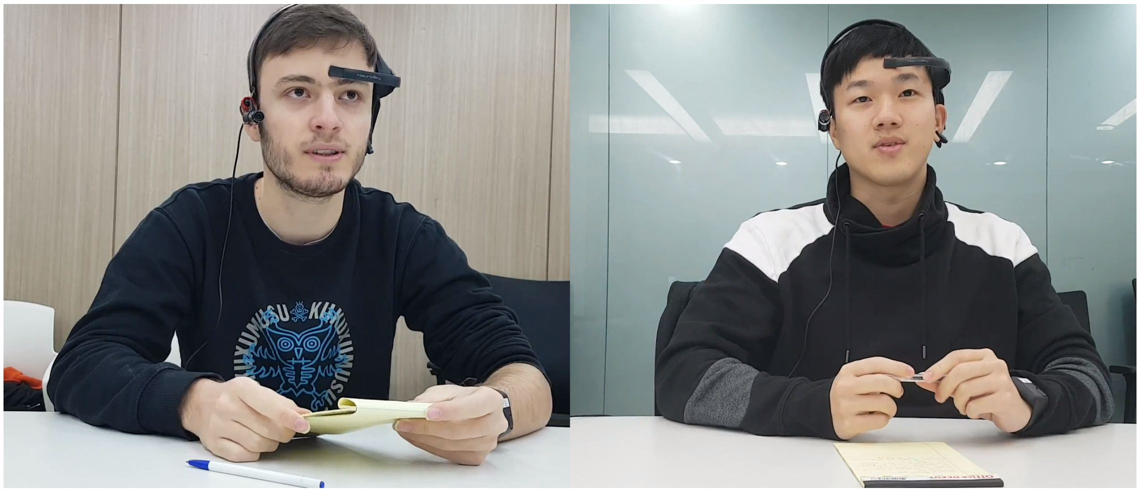


Figure 3.2: A screen captured from actual debate footage of two participants (left: P5 and right: P6).

K-EmoCon is a publicly available multimodal dataset acquired from 32 participants engaging in 10-minute long paired debates on a social issue. It enables studying emotions in the context of naturalistic conversations taking place in a social setting and recognizing continuous emotional states from physiological signals acquired with commercial-grade wearable devices. For data collection, participants were randomly assigned into pairs and engaged in a face-to-face debate on the Jeju Yemeni refugee crisis for approximately 10 minutes. All debate sessions were conducted in a room with controlled temperature and illumination to minimize environmental variations across debates. As shown in figures 3.1 and 3.2, two participants sat across a table facing each other with cameras in the middle, recording their facial expressions and upper body movements, along with speeches during a debate.

Participants wore a suite of mobile, wearable sensing devices during debates for the collection of physiological signals. The choice of commercially available devices makes the dataset particularly fitting for our purpose of developing a sampling method for emotions in the wild compared to many other emotion datasets collected in a controlled laboratory setting with medical-grade devices. See table 3.1 for a detailed summary of the differences between the K-EmoCon and other datasets. Figure 3.3 and table 3.2 show detailed info on the data collection apparatus used in the construction of the K-EmoCon dataset, including how the devices were worn, their sampling rates, and the range of collected signals.

All debates were moderated by an experimenter and lasted approximately 10 minutes. Each participant took turns speaking up to two consecutive minutes, and the moderator stopped a debate at the ten-minute mark with some flexibility. After debates, participants annotated their own emotions and their debate partner's emotions at the interval of every 5 seconds from the beginning to the end of a debate, respectively watching the footage of themselves and their partners during the debate. Additionally, five external raters were recruited to annotate participants' emotions during debates from an external observer's point-of-view.

Table 3.1: Comparison of the K-EmoCon dataset with the existing multimodal emotion recognition datasets: Posed emotions are when a subject is instructed to enact a particular emotion while Spon. = spontaneous. Similarly, induced emotions are when a set of selected stimuli is used for their elicitation. For annotation types, S = *self*, P = *partner*, and E = *external observer*.

Name (year)	**Size**	**Modalities**	**Spon. vs. posed**	**Natural vs. induced**	**Annotation method**	**Annotation type**	**Context**
IEMOCAP (2008) [15]	10	Videos, face motion capture, gesture, speech (audio & transcribed)	Both	Both†	Per dialog turn	S, E	Dyadic
SEMAINE (2011) [90]	150	Videos, FAUs, speech (audio & transcribed)	Spon.	Induced	Trace-style continuous	E	Dyadic
MAHNOB-HCI (2011) [134]	27	Videos (face and body), eye gaze, audio, biosignals (EEG, GSR, ECG, respiration, skin temp.)	Spon.	Induced	Per stimuli	S	Individual
DEAP (2012) [74]	32	Face videos, biosignals (EEG, GSR, BVP, respiration, skin temp., EMG & EOG)	Spon.	Induced	Per stimuli	S	Individual
DECAF (2015) [1]	30	NIR face videos, biosignals (MEG, hEOG, ECG, tEMG)	Spon.	Induced	Per stimuli	S	Individual
ASCERTAIN (2016) [136]	58	Facial motion units (EMO), biosignals (ECG, GSR, EEG)	Spon.	Induced	Per stimuli	S	Individual
MSP-IMPROV (2016) [17]	12	Face videos, speech audio	Both	Both†	Per dialog turn	E	Dyadic
DREAMER (2017) [65]	23	Biosignals (EEG, ECG)	Spon.	Induced	Per stimuli	S	Individual
AMIGOS (2018) [30]	40	Vidoes (face & body), biosignals (EEG, ECG, GSR)	Spon.	Induced	Per stimuli	S, E	Individual, Group
MELD (2019) [114]	7	Videos, speech (audio & transcribed)	Both	Both†	Turn-based	E	Dyadic, Group
CASE (2019) [131]	30	Biosignals (ECG, respiration, BVP, GSR, skin temp., EMG)	Spon.	Induced	Trace-style continuous	S	Individual
CLAS (2020) [85]	64	Biosignals (ECG, PPG, EDA), accelerometer	Spon.	Induced	Per stimuli/task	Predefined‡	Individual
K-EmoCon (2020) [105]	*32*	*Videos (face, gesture), speech audio, accelerometer, biosignals (EEG, ECG, BVP, EDA, skin temp.)*	*Spon.*	*Natural*	*Interval-based continuous*	*S, P, E*	*Dyadic*

† A dataset was considered to contain induced emotions if scripted interaction was involved in the data collection, even though no artificial stimuli (such as an emotion inducing video clip) was used.

‡ Predefined emotion categories of stimuli and success rates of participants in a set of purposefully selected cognitive tasks were used as ground-truth labels.

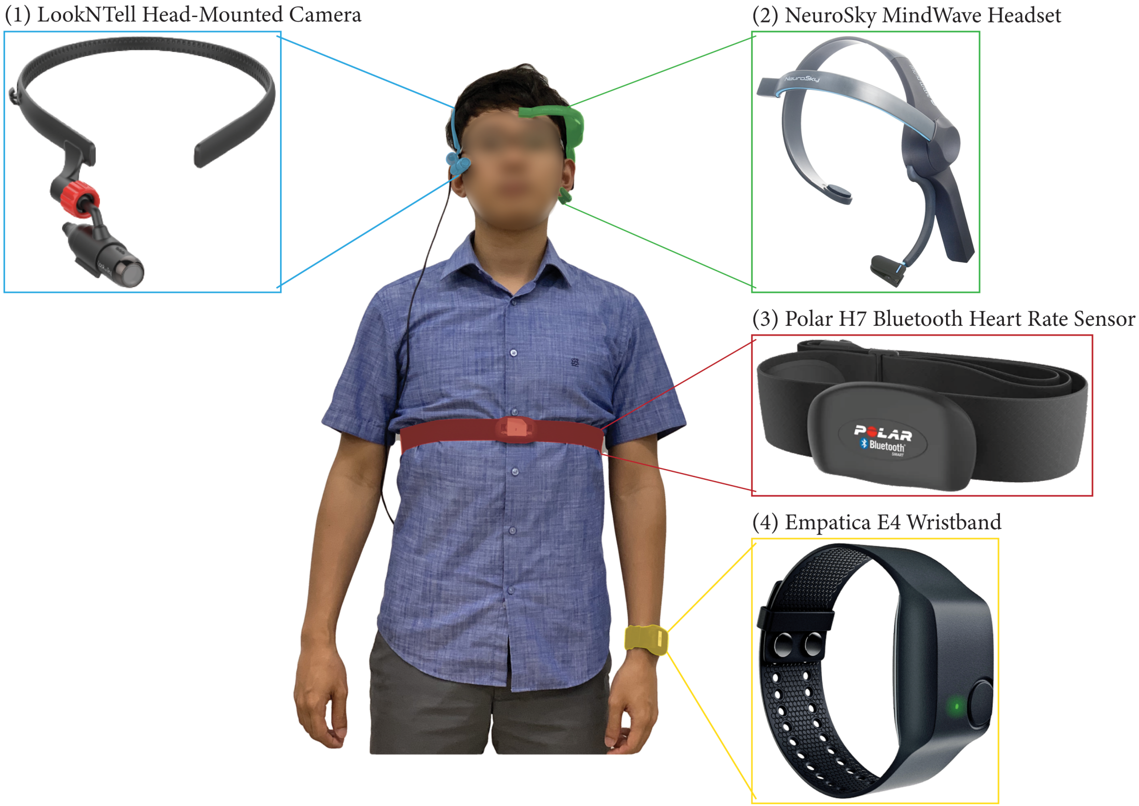

(1) LookNTell Head-Mounted Camera
(2) NeuroSky MindWave Headset
(3) Polar H7 Bluetooth Heart Rate Sensor
(4) Empatica E4 Wristband

Figure 3.3: Wearable sensors worn by participants during data collection sessions for the construction of the K-EmoCon dataset.

Table 3.2: Physiological signals collected with wearable sensing devices in K-EmoCon, with respective sampling rates and signal ranges.

Devices	**Collected data**	**Sampling rate**	**Range [min, max]**
Empatica E4	3-axis acceleration	32Hz	[-2g, 2g]
	BVP (PPG)	64Hz	n/a
	EDA	4Hz	[0.01μS, 100μS]
	Heart rate (from BVP)	1Hz	n/a
	IBI (from BVP)	n/a	n/a
	Body temperature	4Hz	[$-40\,^{\circ}$C, $115\,^{\circ}$C]
NeuroSky MindWave	Brainwave (EEG Fp1)	125Hz	n/a
	Attention & Meditation	1Hz	[0, 100]
Polar H7	HR (ECG)	1Hz	n/a

3.1.2 Dataset contents

The resulting dataset includes data from 16 paired debates, which sum to 172.92 minutes, including physiological signals, audiovisual recordings of debates, and continuous annotations of emotions from three distinct perspectives of the subject, the partner, and the external observers. Table 3.3 summarizes the contents of the dataset.

Table 3.3: Summary of data collection and the contents of the K-EmoCon dataset.

Data collection summary	
Number of participants	32 (20 males and 12 females)
Participants' age	19 to 36 (mean = 23.8 years, stdev. = 3.3 years)
Session duration	Total 172.92 min, (mean = 10.8 min, stdev. = 1.04 min)
Annotated emotions	**1 - 5**: Arousal, Valence **1 - 4**: Cheerful, Happy, Angry, Nervous, Sad **Choose one**: Common BROMP affective categories + less common BROMP affective categories [100]
Collected biosignals	3-axis Acc. (32Hz), BVP (64Hz), EDA (4Hz), heart rate (1Hz), IBI (n/a), body temperature (4Hz), EEG (8 band, 32Hz), ECG (2Hz)

Among the collected data, we focus on physiological signals for the recognition of self-reported emotions. In particular, we use data collected with Empatica E4 (BVP, EDA, and HST) and Polar H7 (ECG) to predict arousal and valence measured with the 5-point Likert scale. This is as our goal is to develop an adaptive sampling mechanism that can trigger a self-report questionnaire at an appropriate moment to sample users' emotions by monitoring their physiological signals.

3.2 Data Preprocessing

Preprocessing raw physiological data and extracting features capturing characteristics and relationships between different types of physiological signals is at the core of developing an emotion recognition system. In the following section, we discuss the procedures in preparing physiological signals and emotion annotations in the K-EmoCon for the training of emotion recognition models.

3.2.1 Feature extraction from physiological signals

For extracting features from physiological signals, we use PyTEAP [104], a Python implementation of Toolbox for Emotion Analysis using Physiological signals (TEAP) [135, 144]. TEAP is a toolbox dedicated to the extraction of features from multiple types of physiological signals, including electrocardiogram (ECG), blood volume pulse (BVP), galvanic skin response (GSR), human skin temperature (HST), electromyogram (EMG), respiration (RES), and electroencephalogram (EEG). While TEAP is written in MATLAB, we translated TEAP into Python, as it is interoperable with popular machine learning and deep learning frameworks such as PyTorch, and created PyTEAP.

Table 3.4: Features extracted from physiological signals with PyTEAP and their descriptions.

Type	Feature	Description
GSR	Peaks per second	average number of peaks in resistance exceeding 100 ohms per second
	Mean amplitude	mean amplitude of peaks from the saddle point preceding the peak
	Mean rise time	the average time for GSR to reach peaks from saddle points in seconds
	Statistical moments	mean and SD
BVP	Interbeat interval (IBI)	mean IBI and HRV (SD)
	Multiscale entropy (MSE)	MSE at 5 levels
	Tachogram power	$\log(P_x^{LF}(f)), \log(P_x^{MF}(f)), \log(P_x^{HF}(f)), \frac{\log(P_x^{MF}(f))}{\log(P_x^{LF}(f))+\log(P_x^{HF}(f))}$
	Power spectral density (PSD)	$\log(P_x^{LF}(f)), \frac{\log(P_x^{LF}(f))}{\log(P_x^{HF}(f))}$
	Statistical moments	mean
ECG (BPM)	Statistical moments	mean and SD
HST	PSD	$\log(P_x(f))$
	Statistical moments	mean, SD, kurtosis, and skew

Compared to TEAP, PyTEAP supports feature extraction for ECG, BVP, GSR, and HSR, which are the types of physiological signals present in the K-EmoCon. Table 3.4 summarizes features supported by PyTEAP. The minimum length of signals PyTEAP supports for feature extraction is 25 seconds. PyTEAP also automatically applies low-pass filters to remove high-frequency perturbations in signals collected at sufficiently high sampling rates. For K-EmoCon, however, filtering is applied only for BVP, which has a sampling rate of 64Hz, at 1/8th of the original rate; other signals are used as is given their sampling rates are already too low for any filtering.

Galvanic skin response – GSR

Galvanic skin response or electrodermal activity is a measure of electrical resistance (or conductivity) on the skin surface [9, 12]. This value can change with the amount of sweat excretion [6, 62], which is under the control of the sympathetic nervous system mediating fight-or-flight response. This indicates that by measuring GSR, one could detect affective states associated with fight-or-flight responses, such as fear and stress. GSR is then characterized into two types of changes — tonic skin conductance level and phasic skin conductance response, each characterizing slowly varying levels of skin conductivity over time without any particular interference from external stimuli and rapid variations in conductance under the effect of some external factor, with peaks referred to as skin conductance responses (SCRs). Thus, our GSR features would be more informative for understanding the tonic changes as they are averaged over a time window greater than 25 seconds, while peaks per second, mean amplitude, and mean rise time may carry some information regarding phasic changes as well.

Blood volume pulse – BVP

The BVP is measured with the photoplethysmography (PPG) and measures changes in blood volume. It is closely related to the interbeat interval (IBI), which measures the time distances between two consecutive heartbeats. Our BVP features measure various characteristics of this signal: mean IBI and the heart rate variability as a standard deviation of IBI, multiscale entropy at 5 levels, tachogram power at low frequencies ($LF : f \in [0.01, 0.08]$Hz), middle frequencies ($MF : f \in [0.08, 0.15]$Hz), high frequencies ($HF : f \in [0.15, 0.4]$Hz), and the ratio between the power of medium to low plus high frequencies, and PSD at four frequency bands of $f \in \{[0, 0.1], [0.1, 0.2], [0.2, 0.3], [0.3, 0.4]\}$Hz, and the ratio between low to high frequencies with LF between $[0.08, 0.15]$Hz and HF between $[0.15, 0.4]$Hz.

Electrocardiogram – ECG

In the K-EmoCon dataset, ECG is equivalent to beats per minute (BPM), and its statistical moments (mean and SD) are used as features.

Human skin temperature – HST

HST is as straightforward as a measure of skin temperature in degrees Celsius. Its PSD are extracted as features similar to BVP, for frequencies in $f \in \{[0, 0.1], [0.1, 0.2]\}$Hz, along with statistical moments including mean, standard deviation, kurtosis, and skew.

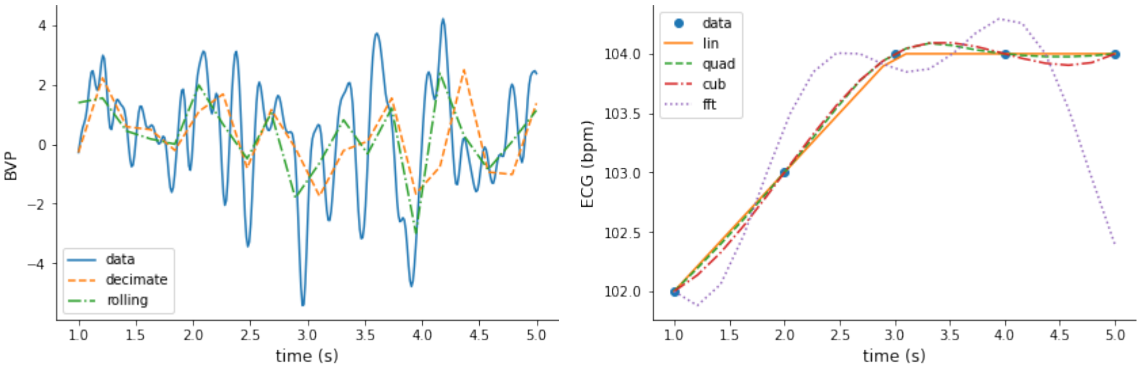
3.2.2 Deep networks for physiology-based emotion recognition

One major difference and advantage of deep neural networks compared to traditional machine learning models is that they allow bypassing feature extraction steps. It is a well-established knowledge that deep networks themselves are automatic feature extractors, i.e., representation learners [10]. For example, convolution neural networks (CNNs) are known to learn simple shapes such as lines and curves in the lower layers and more complex shapes resembling real-world objects in the upper layers [152]. Autoencoders [57] are another example where the feature extraction capability of a DNN is utilized to a full extent. Autoencoders learn to reduce raw data in a higher dimension to a simpler latent vector in a lower dimension by repeating the encoding-decoding process. Also, in natural language processing (NLP), with sufficient data and carefully designed algorithms, human effort in engineering meanings to words to help machines understand human language can be minimized [26, 92].

This notion of using deep networks as automatic extractors of characteristics and relationships embedded in raw data similarly extends to physiological signals [86]. Given that, this work also explores modeling emotions manifested via human physiology with deep neural networks. We utilize a *recurrent neural network* (RNN) [118] known for its ability to learn long-term dependencies from a sequence of inputs changing over time, such as human languages, as physiological signals are also 1-dimensional temporal sequences. In particular, we use a popular variant of RNN, a *long short-term memory* (LSTM) network [58], which is robust against the vanishing gradient problem in traditional RNNs.

We apply minimal preprocessing to physiological signals to prepare K-EmoCon raw data as inputs to a deep neural network. As the sampling rates differ across physiological signals, we first resample BVP and ECG to match their sampling rates to EDA and HST collected at 4Hz.

As shown in figure 3.4, BVP is downsampled from 64Hz to 4Hz, and ECG is upsampled from 1Hz to 4Hz. The decimation function in the SciPy package with a downsampling factor of 16 is used for downsampling BVP, and a quadratic 1-D interpolation function in the same package is used for upsampling ECG. Both functions were chosen heuristically after visually inspecting the shape of resulting signals from different sampling methods. After resampling, four 1-dimensional signals are stacked to form a 2-dimensional array of shape 4×20 corresponding to a 5-second segment of physiological data.



(a) Downsampling BVP

(b) Upsampling ECG

Figure 3.4: Resampling BVP (64Hz) and ECG (1Hz) to 4Hz to match lengths of physiological signals as inputs to an LSTM network.

3.3 Experiment Setup

This section discusses the formalization of emotion recognition using the K-EmoCon dataset as a binary classification problem. Online active learning algorithms for adaptive sampling of emotions in a naturalistic setting and our approach in combining uncertainty sampling and minority sampling strategies into a single parametric query function to enable selective sampling of emotions are proposed. Finally, we discuss how we will evaluate our proposed adaptive sampling method for emotions.

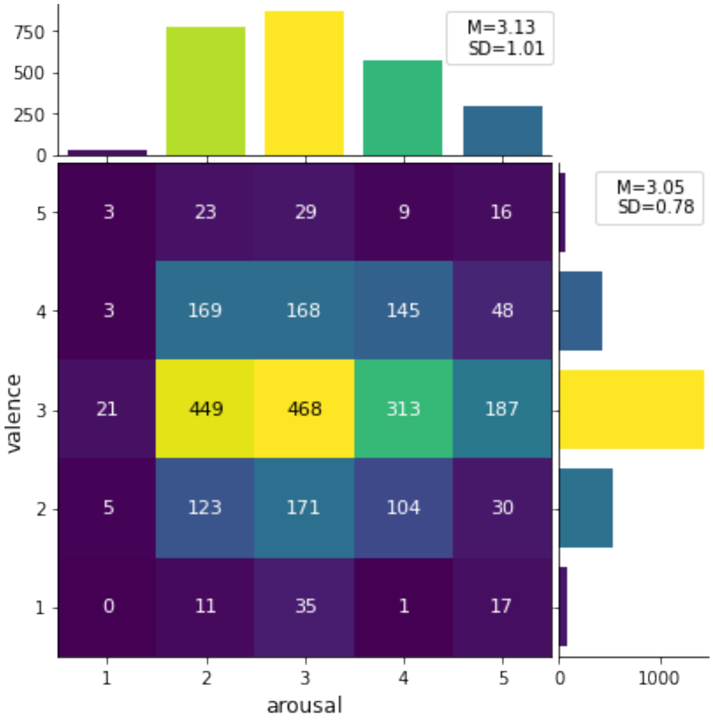


Figure 3.5: The distribution of self-reported arousal and valence labels in the K-EmoCon dataset.

3.3.1 Binary classification of emotions

Emotion recognition in this work is formulated as a binary classification problem between low and high classes, respectively, for two dimensions of arousal and valence. While arousal and valence were both measured on the 5-point Likert scale during the dataset's construction, low and high classes are defined separately for two emotion vectors. For arousal, low class corresponds to {1, 2, 3} including neutral (=3) and high class to {4, 5}, while low class includes {1, 2} and high class {3, 4, 5} for valence. This choice is to avoid conflating negative emotions with a neutral state.

As in figure 3.5, the self-reports of arousal and valence in the K-EmoCon dataset are centered around neutral with a slight bias towards the positive side. This is unsurprising as negative emotions in the wild are reported less frequently compared to positive emotions. While people default to a neutral state in general, social desirability bias [20] and people's positive reappraisal of past events as a coping mechanism [27, 79] contribute to the imbalance in the distribution of self-reported emotions. Nonetheless, correct recognition of negative emotions intuitively has more pressing implications than recognizing positive emotions as they can signal potential mental illness or social conflicts. Given that, in this work, the neutral state is grouped with low arousal (LA) and high valence (HV), to make high arousal (HA) and low valence (LV) states associated with negative emotions more pronounced.

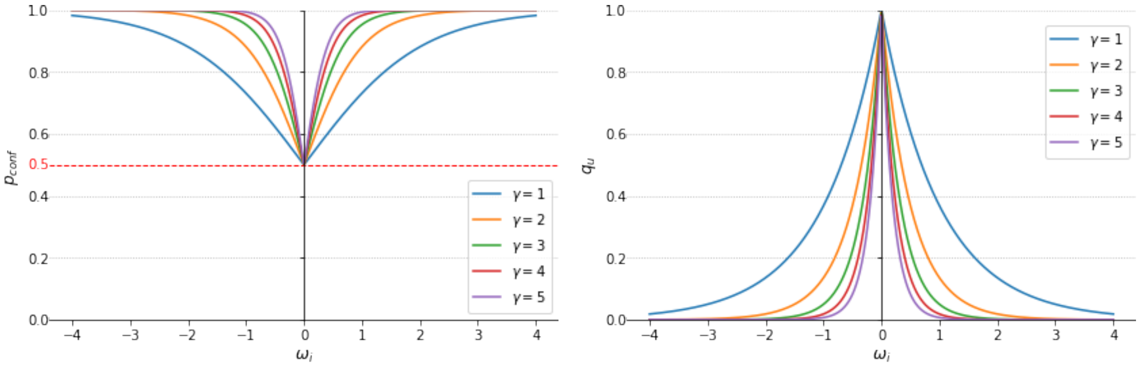
3.3.2 Query strategy for active learning

In the following, an instance in the dataset $\mathcal{K}$ is a pair (x_i, y_i), where x_i is an input segment of physiological signals of varying length, starting from the minimum of 5 seconds, and y_i is a corresponding emotion label for either valence or arousal, which takes on the values {0, 1}, representing low or high.

An active learner is then assumed to be constructed upon a classification model $\omega_i = f(x;\theta)$. A model, defined by parameters θ, takes an input x_i in a set X of size M and returns ω_i in a model's sample space Ω of size N with $M \leq N$. A model output ω_i is a single value representing the distance of the input x_i from a classification hyperplane. A value of ω_i below zero indicates that a model predicts x_i as low, and vice versa for a value above zero, while a large absolute value of ω_i signals that a model has high confidence in its prediction. To convert this ω_i into a binary prediction for emotion label $\hat{y}_i$, we define a `classify` function that takes ω_i as an input and returns a predicted class label $\hat{y}_i$ associated with an input x_i:

$$\hat{y}_i = \texttt{classify}(\omega_i) = \begin{cases} 0 & \text{if } \omega_i < 0 \\ 1 & \text{else if } \omega_i > 0 \\ \sim \text{Bernoulli}(0.5) & \text{otherwise } \omega_i = 0 \end{cases} \tag{3.1}$$

Note that when ω_i equals zero, a model is uncertain of its prediction, and the given input is equally likely to be classified as low or high. Therefore, we require a model to query the associated label on such occasions and use the newly acquired information in future predictions. This behavior is a defining characteristic of an active learner. For the rest of the paper, a "learner" and a "model" all refer to an active learner capable of actively querying and learning incrementally, and we will use the terms interchangeably.



(a) Confidence

(b) Uncertainty

Figure 3.6: Query probability functions for uncertainty sampling.

Uncertainty sampling

The approach of querying for examples that a learner is least sure about is *uncertainty sampling*. Specifically, we employ a *logistic margin sampling*, which was previously used in activity recognition [2, 93] and spam mail filtering [127]. For logistic margin sampling, we first compute the confidence of a learner p_{conf} given an input x_i using a sigmoid function, then use p_{conf} to obtain the probability q_u that a learner would query for the emotion label y_i associated with the input:

$$p_{\text{conf}} : \mathbb{R} \to [0.5, 1] : \omega_i, \gamma \mapsto \frac{1}{1 + e^{-\gamma|\omega_i|}} \tag{3.2}$$

and

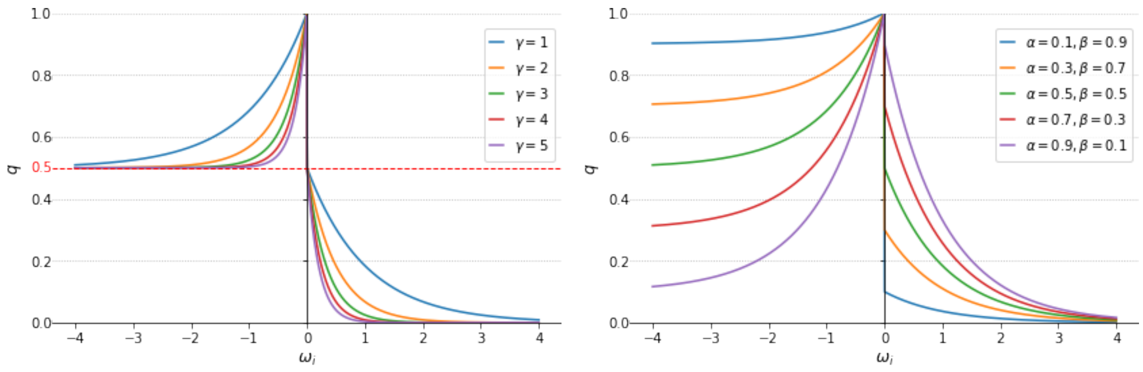
$$q_u : \mathbb{R} \to [0, 1] : p_{\text{conf}} \mapsto \frac{1}{p_{\text{conf}}} - 1 \tag{3.3}$$

Here, p_{conf} is a probability that a learner expects the predicted label $\hat{y}_i$ to match the ground truth. Of course, as $p_{\text{conf}} = e^{-\gamma|\omega_i|}$, it is possible to bypass calculating p_{conf} and directly get q_u. An absolute value is taken for ω_i as p_{conf} can represent a learner's confidence in both negative and positive directions; thus, a learner is least confident when p_{conf} equals 0.5, meaning that a learner is only confident as much as a random classifier. The additional parameter γ, which can be any non-negative number including zero, then controls a learner's behavior, such that a learner is more confident in its prediction with a higher γ.

As in figure 3.6a, while p_{conf} always equals 0.5 when $\omega_i = 0$, p_{conf} gets higher for greater values of γ while ω_i stays the same. This can be interpreted as that a learner is more selective in its queries with higher γ values. Figure 3.6b shows plots of q_u at different values of γ. Like p_{conf}, a learner will always query with $q_u = 1$ at ω_i, but at other values of ω_i, a learner will increasingly query only for instances it finds near the hyperplane as γ increases.

Minority sampling

Another query strategy in active learning is *minority sampling* [82], where a learner actively queries labels for inputs that it expects to belong to a minority class. This concept has not been explored to the extent of uncertainty sampling in the literature. However, the idea is similarly introduced in other works as diversity sampling [130], where the sample space is divided into groups such that instances in a group



(a) Query probability at different values of γ

(b) Query probability at varying levels of α and β

Figure 3.7: Query probability functions for paremeterized uncertainty-minority sampling.

are similar to each other than instances in other groups, and instances are evenly sampled from each group. This strategy aims to make the distribution of the dataset resulting from data collection balanced and a resulting model to make better predictions/classifications for samples that belong to the minority class. This strategy is appropriate in our scenario of collecting emotions in the wild, as the distribution of naturally occurring emotions is biased towards positive, while the recognition of negative emotions is often more critical in applications based on emotion recognition such as depression prognosis.

With the assumption minority class label = 0, minority oversampling for an active learner can be defined as a simple conditional function:

$$q_m = \begin{cases} 0 & \text{if } \texttt{classify}(\omega_i) = 1 \\ 1 & \text{else if } \texttt{classify}(\omega_i) = 0 \end{cases} \tag{3.4}$$

Note that the minority class label can change as an active learner accumulates more samples throughout training, and the distribution of the collected samples changes. In that case, a learner should update what it considers as a minority class. However, there can be a problem if the distribution of the samples queried during the learning process and the distribution samples to be predicted differ significantly, as then a learner would be overfitting to the training data and fail to estimate the target distribution, rendering itself useless in realistic scenarios.

Parameterized uncertainty-minority sampling

To overcome the limitations of only emphasizing minority samples, we can combine the previous two query strategies into a single strategy, a *parameterized uncertainty-minority sampling*, and take advantage of both the uncertainty and minority sampling. The proposed sampling strategy is as follows:

$$\begin{gathered} \texttt{query}(\omega_i) = \text{Bernoulli}(\alpha \cdot q_u(\omega_i) + \beta \cdot q_m(\omega_i)) \\ \text{for } \alpha + \beta = 1 \text{ and } 0 \leq \alpha, \beta \leq 1 \end{gathered} \tag{3.5}$$

This query function, parameterized with three parameters α, β, and γ, and taking a model output ω_i as an input, returns a value between 0 and 1, which is the probability a learner would query for the label associated with the input, mapping model outputs ω_i to query probabilities.

While γ controls a learner's selectivity, as in figure 3.7a, α and β control the respective influence of uncertainty sampling (α) and minority oversampling (β) in decisions to query or not. Figure 3.7b shows as α trades off with β, i.e., more weight is put on querying samples a learner is uncertain with than querying minorities, the shape of the query function approaches the shape of the uncertainty sampling query function as in figure 3.6b.

3.3.3 Online active learning

Two major learning scenarios in the active learning framework are *pool-based active learning* and *stream-based selective sampling* (or stream-based active learning). This work focuses on stream-based active learning, assuming the *online learning* [129] setting where a model decides to query or discard items from some stream of input sources one by one and learns on the fly.

Single-pass stream-based active learning

In particular, we define a sequence of steps where active learning occurs given a dataset consisting of inputs and associated labels as *single-pass stream-based active learning*. Algorithm 1 illustrates the proposed approach.

Algorithm 1: Single-pass Stream-based Active Learning

```
Data: $\mathcal{K} \subseteq \mathbb{R}^{N \times D}$
Input: init_size, test_size, update_size, learning_rate: $\eta$, query_params: $\{\alpha, \beta, \gamma\}$
initialize by splitting the dataset $\mathcal{K}$ into:
    1. initial train set $\mathcal{R} = (X_\mathcal{R}, Y_\mathcal{R})$,
    2. test set $\mathcal{T} = (X_\mathcal{T}, Y_\mathcal{T})$,
    3. data stream $\mathcal{S} = (X_\mathcal{S}, Y_\mathcal{S})$,
such that $|\mathcal{R}| =$ init_size, $|\mathcal{T}| =$ test_size, and $|\mathcal{S}| = N -$ init_size $-$ test_size;
initialize the active learner $f_\theta : X \to Y$, w/ random parameters $\theta$;
fit $f_\theta$ to $\mathcal{R}$, with the update rule $\theta \leftarrow \theta - \eta \nabla \mathcal{L}(f_\theta(X_\mathcal{R}), Y_\mathcal{R})$;
initialize the query buffer $\mathcal{Q} \leftarrow [\emptyset]$;
for $x_\mathcal{S}^{(i)} \in \mathcal{S}$, where $i = 1, 2, ..., |\mathcal{S}|$ do
    get $\omega_i = f_\theta(x_\mathcal{S}^{(i)})$;
    if query($\omega_i$) then
        update $\mathcal{Q} \leftarrow \mathcal{Q} + (x_\mathcal{S}^{(i)}, y_\mathcal{S}^{(i)})$;
    end
    if $|Q| \geq$ update_size then
        update $\mathcal{R}$, where $X_\mathcal{R} \leftarrow X_\mathcal{R} + X_\mathcal{Q}$ and $Y_\mathcal{R} \leftarrow Y_\mathcal{R} + Y_\mathcal{Q}$;
        incrementally fit $f_\theta$ to $\mathcal{R}$, with $\theta \leftarrow \theta - \eta \nabla \mathcal{L}(f_\theta(X_\mathcal{R}), Y_\mathcal{R})$;
        test updated model model $f_\theta$ with $\mathcal{T}$;
        empty query buffer $\mathcal{Q} \leftarrow [\emptyset]$;
    end
end
test the final model $f_\theta$ with $\mathcal{T}$;
```

The decision to query or not is made only once for each sample in this algorithm. This single-pass approach is particularly appropriate in the context of sampling emotions in the wild, where emotions in retrospect are subject to numerous biases [5], including memory/recall bias and self-report biases.

Note that although the decision is made only once for each sample, samples are used multiple times for updating parameters during the training process.

Model coverage

However, a limitation in this approach is that it cannot control the *coverage* of a learner, i.e., how much of the total data should an active learner query. To compensate for this issue, we refine algorithm 1 by modifying the query decision rule to consider the model's coverage. For that, we define an *empirical coverage* ϕ_m [49] as the following, with $|\mathcal{R}|$ denoting the size of the initial training batch:

$$\phi_m = \frac{|\mathcal{R}| + \sum_{i=1}^{m} \texttt{query}(\omega_i)}{|\mathcal{R}| + m} \tag{3.6}$$

Algorithm 2: Single-pass Stream-based Active Learning with Coverage

```
Data: $\mathcal{K} \subseteq \mathbb{R}^{N \times D}$
Input: init_size, test_size, update_size, learning_rate: $\eta$, query_params: $\{\alpha, \beta, \gamma\}$,
       target_coverage: $\phi^*$
initialize by splitting the dataset $\mathcal{K}$ into:
   1. initial train set $\mathcal{R} = (X_\mathcal{R}, Y_\mathcal{R})$,
   2. test set $\mathcal{T} = (X_\mathcal{T}, Y_\mathcal{T})$,
   3. data stream $\mathcal{S} = (X_\mathcal{S}, Y_\mathcal{S})$,
such that $|\mathcal{R}| =$ init_size, $|\mathcal{T}| =$ test_size, and $|\mathcal{S}| = N -$ init_size $-$ test_size;
initialize the active learner $f_\theta : X \to Y$, w/ random parameters $\theta$;
fit $f_\theta$ to $\mathcal{R}$, with the update rule: $\theta \leftarrow \theta - \eta \nabla \mathcal{L}(f_\theta(X_\mathcal{R}), Y_\mathcal{R})$;
initialize coverage: $\phi_m \leftarrow 1$, # queried samples: $m \leftarrow |\mathcal{R}|$, and the query buffer: $\mathcal{Q} \leftarrow [\emptyset]$;
for $x_\mathcal{S}^{(i)} \in \mathcal{S}$, where $i = 1, 2, ..., |\mathcal{S}|$ do
    get $\omega_i = f_\theta(x_\mathcal{S}^{(i)})$;
    if $\phi_m < \phi^*$ or query($\omega_i$) then
        update $\mathcal{Q} \leftarrow \mathcal{Q} + (x_\mathcal{S}^{(i)}, y_\mathcal{S}^{(i)})$;
        increment $m \leftarrow m + 1$;
    end
    update $\phi_m \leftarrow \frac{m}{\texttt{init\_size}+i}$;
    if $|Q| \geq$ update_size then
        update $\mathcal{R}$, where $X_\mathcal{R} \leftarrow X_\mathcal{R} + X_\mathcal{Q}$ and $Y_\mathcal{R} \leftarrow Y_\mathcal{R} + Y_\mathcal{Q}$;
        incrementally fit $f_\theta$ to $\mathcal{R}$, with $\theta \leftarrow \theta - \eta \nabla \mathcal{L}(f_\theta(X_\mathcal{R}), Y_\mathcal{R})$;
        test updated model model $f_\theta$ with $\mathcal{T}$;
        empty query buffer $\mathcal{Q} \leftarrow [\emptyset]$;
    end
end
test the final model $f_\theta$ with $\mathcal{T}$;
```

By definition, ϕ_m is the ratio of samples in a full dataset a learner has labels for at a given point in time when m samples from a stream have been inspected for querying. If this value is below a threshold, we may tweak a learner's behavior to query more samples. While there can be different approaches, we use a heuristic where a learner queries all samples that it encounters when its coverage is below a threshold, otherwise queries using the probabilistic query function discussed above.

While giving an active learner control over how much data it queries, this approach can also benefit incremental learning by mitigating a model's instability during the early learning phase. Similar to the *cold start problem* in recommender systems [80], an active learner may fail to make sufficient or appropriate query requests if it is overfitting to the initial training batch that contains samples disparate from the rest of the dataset. In this case, a learner would make misguided decisions when it encounters new samples. However, with this coverage-based approach, a learner gets to access a portion of data that is at least larger than what is specified by the target coverage. Although the representativeness of samples in the initial batch is still out of control, increasing the dataset size is well known as one of the regularization techniques in machine learning.

3.3.4 Evaluation metrics for imbalanced data

For the evaluation of classification models' performance with imbalanced test data, we use unweighted accuracy as a base metric and two additional metrics of weighted F1 score and weighted AUROC, the area under the receiver operating characteristic (ROC) curve [43].

F1 score is a metric combining precision $= \frac{TP}{TP+FP}$ and recall $= \frac{TP}{TP+FN}$ into one with the harmonic mean is calculated as follows:

$$F_1 = \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3.7}$$

While F1 score is appropriate for binary classification problems, it prioritizes the positive class, thus we use AUROC as another evaluation metric.

AUROC is the area under the ROC curve showing the tradeoff between true positive rate ($TPR = \frac{TP}{TP+FN}$, i.e., recall) and false positive rate($TPR = \frac{TP}{TP+FN}$). The ROC curve is also useful as it allows comparing models with a random classifier. While it is suggested that the precision-recall plot is better than the ROC plot for evaluating binary classification with an imbalanced dataset [123], we opt for AUROC as we would like to have a metric that puts equal importance to positive and negative classes.

Table 4.1: The baseline classification results for arousal and valence with XGBoost at varying window sizes.

	Arousal			Valence		
Metrics	Acc.	F1	AUROC	Acc.	F1	AUROC
25s	0.840	0.836	0.883	0.849	0.831	0.876
30s	0.807	0.801	0.910	0.811	0.786	0.895
35s	0.837	0.836	0.901	0.851	0.835	0.916
40s	0.858	0.855	0.923	0.866	0.956	0.935
45s	0.894	0.893	0.936	0.892	0.890	0.931
50s	0.879	0.877	0.955	**0.945**	**0.943**	**0.980**
55s	0.906	0.904	0.957	0.917	0.914	0.944
60s	**0.919**	**0.918**	**0.958**	0.911	0.908	0.965

Chapter 4. Experiments and Results

This chapter discusses how we conduct experiments to empirically evaluate our proposed active learning approaches for the adaptive sampling of emotions in the wild. In particular, we focus on answering the following questions:

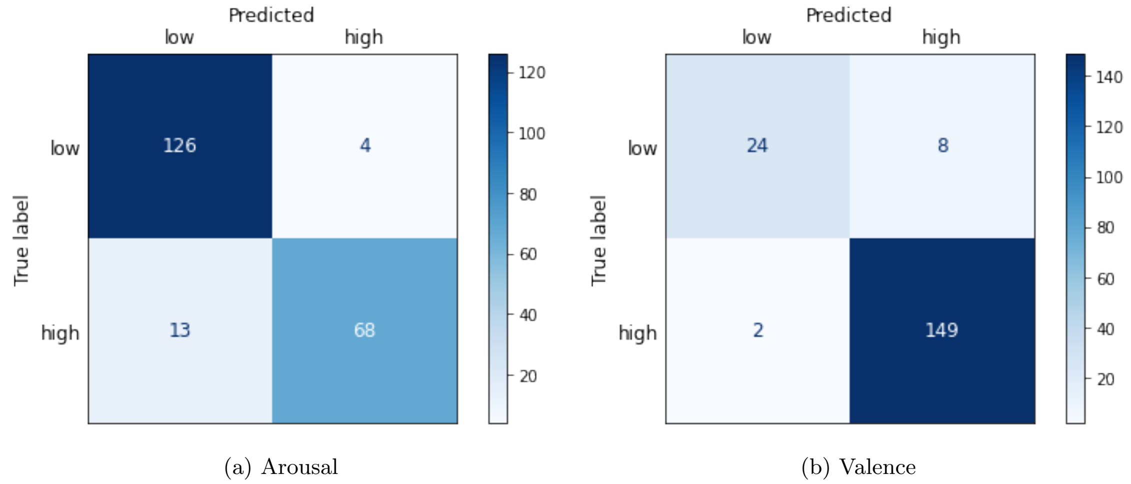
- How does the model's selectivity (γ) affect the performance of an active learner and the size and label distribution of the resulting dataset?
- How does controlling for α and β of the parameterized query function affect the model performance and the resulting dataset?

4.1 Baseline Classification

As discussed in chapter 3, we formulate emotion recognition as a binary classification problem between low/high classes for arousal and valence. However, as the K-EmoCon is a recently published dataset without any previously reported machine learning models trained and evaluated with the dataset, we first train base classification models with K-EmoCon and demonstrate that the K-EmoCon is applicable to classification tasks. We use an XGBoost [24] and an LSTM network for this task, and models are trained separately to predict arousal and valence. Note that we perform holdout cross-validation to evaluate classifiers.

4.1.1 XGBoost

XGBoost is an algorithm based on gradient boosting [47], an ensemble technique where many weak learners are trained additively to form a single robust learner. Each new learner is built greedily upon the residuals of prior learners until no additional improvements are possible. This algorithm applies to both decision trees and linear regressors, and many competitions and research projects employed the algorithm recently for its speed, flexibility, and scalability.



(a) Arousal

(b) Valence

Figure 4.1: Confusion matrices for the baseline classification with XGBoost.

In the following, we use a tree-based XGBoost classifier for both arousal and valence. A classifier is trained at the learning rate of 0.3 with a maximum depth of six for 100 boosting iterations. A total of 30 features extracted from a segment of physiological signals corresponding to the minimum of 25 seconds up to 60 seconds during a debate, containing BVP, EDA, ECG, and HST, are provided as inputs to the classifier. Associated low and high labels are defined separately for arousal and valence, with LA corresponding to {1, 2, 3} including neutral (=3) and HA to {4, 5}, while LV = {1, 2} and HV = {3, 4, 5}. Note that labels for the last 5 seconds in the segments are used as labels for the entire segment. Table 4.1 summarizes the classification results for arousal and valence using an XGBoost with features extracted from varying length of segments from 25 seconds to 60 seconds.

4.1.2 LSTM

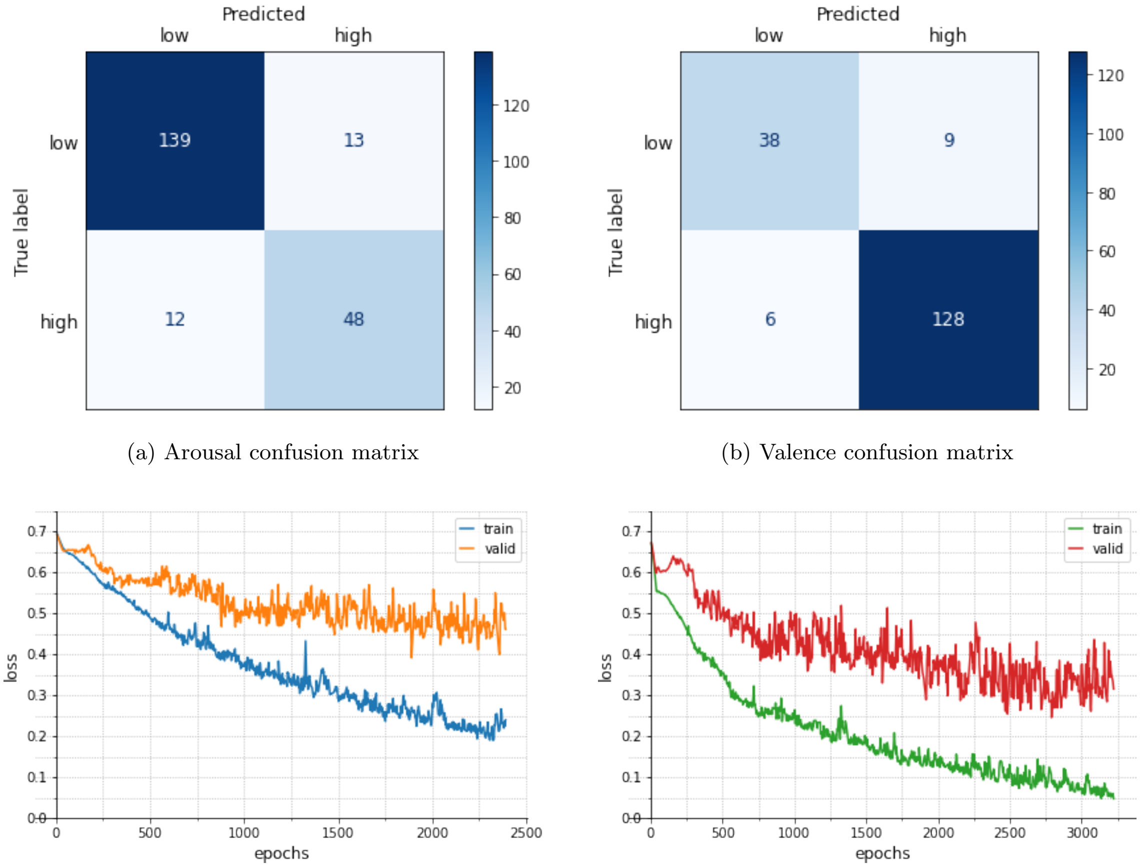
As discussed earlier in section 3.2.2, we also employ LSTM networks for the binary classification of arousal and valence, to verify that K-EmoCon is also usable with deep learning models. We train a 2-layered bidirectional LSTM network with 20 hidden units and a dropout probability of 0.2 between layers, while halting the training early if validation loss does not improve for 500 epochs. For all runs, we use a learning rate of 8.5×10^{-4}. Inputs to the network are resampled BVP, EDA, ECG, and HST signals at 4Hz, which are stacked into 2-D arrays of shape 4 by 4 times the length of segments in seconds, i.e., 25s of physiological signals form an input of size D = 4 by 100, and 4 by 120 from 30 seconds signals, and onward. Arousal and valence labels are defined the same as how we defined them for XGBoost models. Table 4.2 and figure 4.2 shows the results of baseline classification with LSTM networks.

4.2 Simulated Stream-based Active Learning

Given that the K-EmoCon dataset supports classification with both traditional ML models and deep neural networks, we conduct active learning experiments by simulating an online learning scenario with the dataset. In this scenario, the dataset is divided into three parts: 1) initial batch, 2) data stream and 3) test set. An initial batch is for training an initial model that determines whether to query or not for each sample from the data stream. As its name suggests, a data stream simulates a hypothetical

Table 4.2: The baseline classification results for arousal and valence with LSTM at varying input sizes.

	Arousal			Valence		
Metrics	Acc.	F1	AUROC	Acc.	F1	AUROC
25s	0.699	0.694	0.757	0.823	0.820	0.788
30s	0.701	0.692	0.737	0.805	0.797	0.836
35s	0.757	0.748	0.804	0.852	0.845	0.822
40s	0.791	0.790	0.828	0.818	0.815	0.814
45s	0.817	0.814	0.854	0.882	0.884	0.903
50s	0.852	0.854	0.899	0.902	0.903	0.874
55s	0.771	0.766	0.843	0.885	0.885	0.896
60s	**0.882**	**0.865**	**0.936**	**0.917**	**0.916**	**0.942**



(a) Arousal confusion matrix

(b) Valence confusion matrix

(c) Arousal train/validation losses

(d) Valence train/validation losses

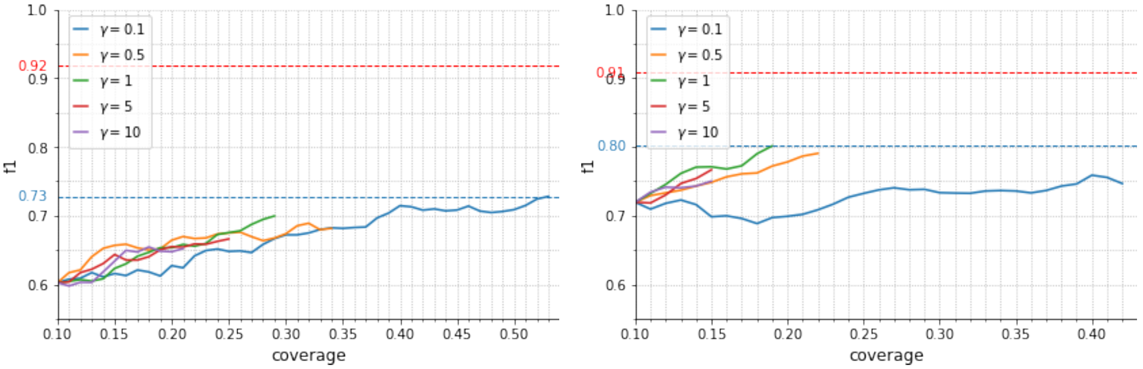
Figure 4.2: Confusion matrices and train/validation losses for the baseline classification with LSTM.

data source that sequentially generates samples for online learning. Upon receiving a sample, a model inspects it and retrieves an associate label if deemed suitable for querying, and saves (sample, label) to a queried samples buffer. A model is updated when the query buffer reaches a predetermined size (e.g., 1% of the full dataset), and the intermediate model's performance is evaluated with the test set. This process continues until there are no more samples left in the data stream.

4.2.1 Query selectivity and model performance

We first observe the effect of the model's selectivity controlled by the γ parameter on the classification performance. As discussed in section 3.3.2, γ affects how the model queries for samples that it finds uncertain in classification. With higher γ, the model will increasingly query for samples that it finds nearer to the hyperplane, and for the same sample, a model will be less likely to query for a label.

To observe how this parameter affects a model's performance, we compare the performance of XGBoost classifiers trained at different levels of γ while holding everything else the same (α and β are both set to 0.5), using the single-pass active learning (algorithm 1) where models are trained incrementally without the target coverage. Figure 4.3 summarizes the results of this experiment:



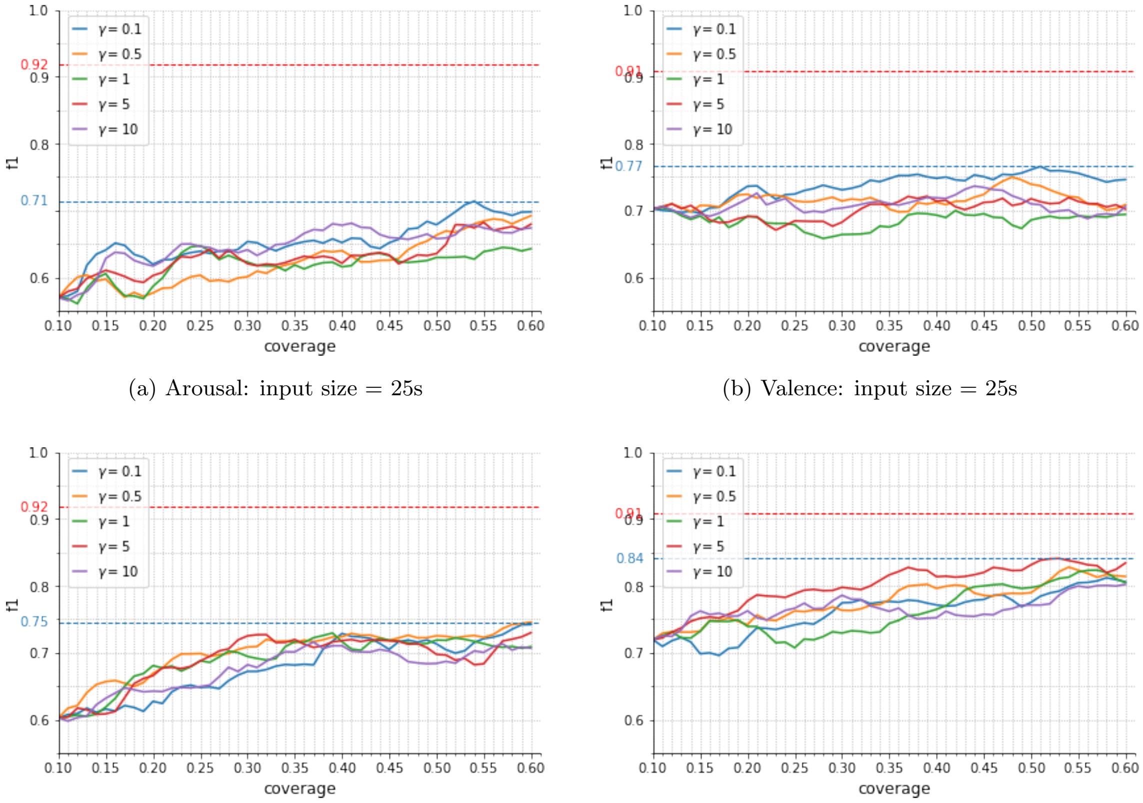
(a) Arousal (b) Valence

Figure 4.3: Learning trajectories of XGBoost-based active learners without target coverage and 60s input features.

Unsurprisingly, the resulting dataset's size is larger at lower γ values; the more data is accumulated overall, the less discriminating a model is in queries. As one can expect, this larger pool of training samples leads to higher model performance for arousal classification. However, this generally accepted tendency between the size of training data and model performance is reversed for valence.

For valence, model performance is the lowest when the γ value is the smallest ($\gamma = 0.1$). This result may, in part, be due to the severe imbalance in valence labels. The positive to the negative ratio for valence is $1357 : 444 = 1 : 0.327$, while arousal is also imbalanced but not as much as valence with the positive to negative ratio of $1346 : 771 = 1 : 0.573$. Given that, we can interpret this result as a low selectivity can benefit the model by allowing the accumulation of a larger set of training samples that, in turn, provides a regularization effect. However, indiscriminate querying can be harmful when the sample distribution is highly biased, and especially when a model is unstable during an early stage of training as it suffers from a cold start. Putting this into parallel, imagine an average student learning to solve math problems from a book that contains some highly challenging, college-level problems (negative samples) and many relatively simple problems at his level (positive samples). If the student were to practice solving problems from this book to be tested later, how can he score higher, practice problems at random throughout the book, or focus on practicing ones among simpler problems that he can somewhat understand but unsure how to solve?

Nonetheless, the effect of γ on model performance is diminished when the target coverage is set.



(a) Arousal: input size = 25s

(b) Valence: input size = 25s

(c) Arousal: input size = 60s

(d) Valence: input size = 60s

Figure 4.4: Learning trajectories of XGBoost-based active learners with target coverage = 60%.

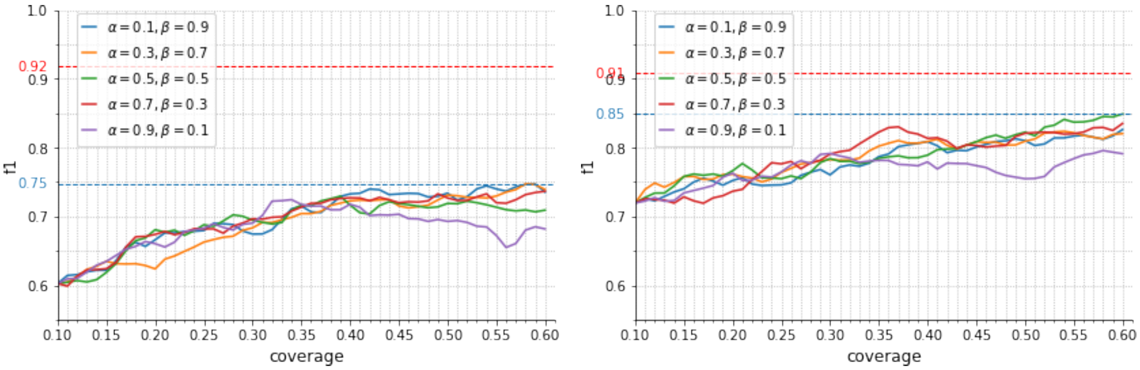
Figure 4.4 shows the result of an experiment where we similarly train XGBoost classifiers for arousal and valence using the single-pass method but using algorithm 2 with the target coverage, which controls the minimum amount of data a model should query. The results show that when the target coverage is set to 60% of the full dataset, while models' performances improve with more training samples, the γ parameter becomes relatively meaningless. Although there are slight differences in each model's learning trajectory, the differences are not significant enough.

One thing we may note, however, is that training with low γ tends to result in better model performance in the long run, based on our empirical observation. This may be specific to the context of emotion recognition, which usually concerns physiological signals collected from individuals with distinct physiology and involves datasets relatively small in size compared to other problems such as activity recognition, where data is abundant and easy to collect. Given these characteristics of scarce and heterogeneous data, the optimum strategy in learning emotions from physiological data might be to take as many chances of exploration whenever possible instead of relying on dubious estimates patched together from a handful of evidence, similar to the notion of exploration and exploitation in reinforcement learning [139].

4.2.2 Tradeoff between uncertainty and minority sampling

In this section, we observe the tradeoff between uncertainty sampling and minority sampling in their effect on model performance. Uncertainty sampling in active learning aims to query instances closest to the decision boundary, which a model is least confident about, and are also the most informative instances likely to cause the most considerable change in the model. Minority sampling, on the other hand, is as straightforward as querying instances that are expected to belong to a minority class, while the definition of the minority group is updated as learning progresses and the distribution of training samples change.

To combine two sampling strategies, we use a parameterized query function defined as $\texttt{query}(\omega_i) = \text{Bernoulli}(\alpha \cdot q_u(\omega_i) + \beta \cdot q_m(\omega_i))$ discussed earlier in section 3.3.2. In this function, α and β are parameters determining the weights to uncertainty and minority sampling, respectively, softly constrained such that they sum to one and are both in the range $[0, 1]$. Using α and β, we experiment with how controlling them affects model performance and the data collection.

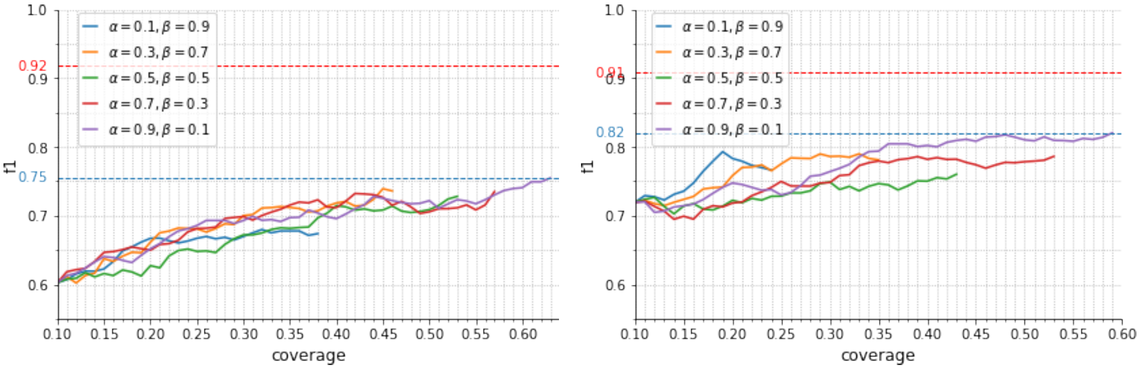


(a) Arousal

(b) Valence

Figure 4.5: Learning trajectories of XGBoost-based active learners at varying levels of α and β, with $\gamma = 1$ and $\phi^* = 0.6$.

First, we observe how different α and β levels affect the performance, while we constrain the model's selectivity to $\gamma = 1$ and collect at least 60% of the total data by setting the target coverage. The results in figure 4.5 show that the classification performance for both arousal and valence is the lowest for $\alpha = 0.9$ and $\beta = 0.1$, i.e., when a larger weight is put on uncertainty sampling than minority sampling. However, a different picture is painted when classifiers are trained at a lower γ without target coverage.



(a) Arousal

(b) Valence

Figure 4.6: Learning trajectories of XGBoost-based active learners at varying levels of α and β, with $\gamma = 0.1$ and no target coverage.

As in figure 4.6, without target coverage and γ set to 0.1, $\alpha = 0.9$ and $\beta = 0.1$ resulted in the best classification performances. Results of these two experiments together point to the conclusion that more exploration can benefit the incremental learning of an active learner in the physiology-based emotion recognition scenario.

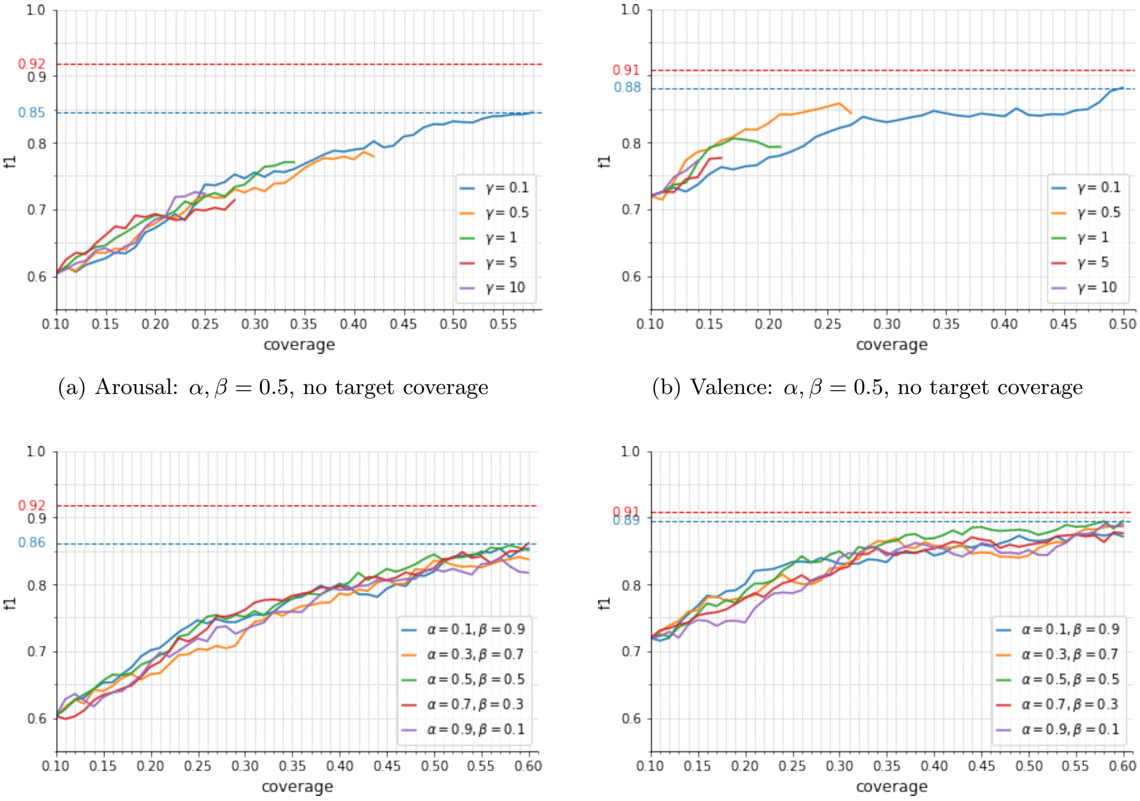
With a target coverage set, a model queries all samples from the stream when the model coverage is below the threshold. If its coverage is over the threshold, then the model decides to query or not following a query function that resembles a step function where $y = 1$ for all x less than or equal to 0, $y = 0$ for all x above 0 (see Fig. 3.7b). Given such a shape of the query function, the model can have more room for exploration with a higher weight on minority sampling than uncertainty sampling. In contrast, without target coverage, a model's final performance is higher with more weights on uncertainty sampling. This can be attributed to the underlying distribution of the K-EmoCon dataset, which is not in favor of minority sampling. As the dataset is inherently imbalanced, it is natural that a model learning from such data more often predicts that a sample belongs to a majority class than a minority class; more queries are likely to be issued with uncertainty sampling than minority sampling when a model is less prone to predict that some sample belongs to a minority group. Thus, the model with a high α will have more chances to explore the sample space with more queries and benefit from the larger pool of training samples, which leads to better model performance.

4.2.3 Rebuilding models per updates

The final performance of active learners so far was subpar to baseline classifiers' performance trained with the full data. Nonetheless, this shortcoming of active learners can be easily overcome by rebuilding the models from scratch after each update of training samples instead of incrementally training existing models.

In this approach, after newly queried samples are added to the existing training set, a new randomly initialized model is trained with the updated train set and replaces the existing model to continue the learning process. Active learning models resulting via this approach show performance comparable to models trained with the full dataset only using 60% of the entire dataset. This result as shown in figure 4.7 suggests that a model can learn better during the online learning process with an artificially induced catastrophic forgetting. Intuitively, this makes sense if we consider our dataset is small, and early models trained from an even smaller portion of the entire dataset is unlikely to have acquired knowledge that is meaningful or will be relevant in making predictions later in the learning process when the amount of training data has substantially increased. So that, forgetting all previously acquired information and starting anew allows an active learner to escape the local minima the earlier model was optimized for and continue learning without being misguided by the wrong knowledge.

This approach will be inapplicable when it comes to larger datasets, especially with deep neural networks. In this case, the model may acquire critical information during the early stages of learning, and being unable to build upon this knowledge may impede the learning process. Of course, once the model could discover the knowledge, then it could rediscover what it forgot, as our XGBoost models did, but this can be challenging when the amount of the data the model has to learn from is significantly larger. Also, for large models trained with sizable data, rebuilding will be simply too inefficient or even impossible when the full data cannot be accessed due to the limited memory.



(a) Arousal: $\alpha, \beta = 0.5$, no target coverage

(b) Valence: $\alpha, \beta = 0.5$, no target coverage

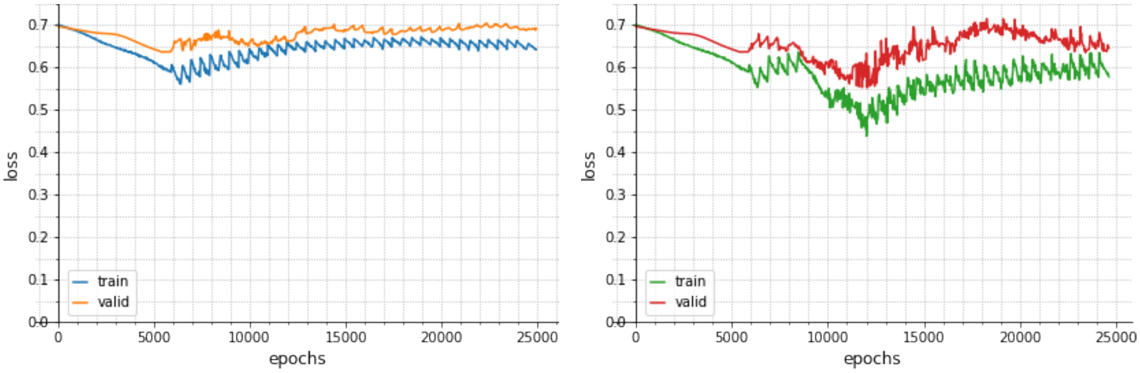
(c) Arousal: $\gamma = 1$, $\phi^* = 0.6$

(d) Valence: $\gamma = 1$, $\phi^* = 0.6$

Figure 4.7: Learning trajectories of XGBoost-based active learners with model rebuilds after each update.

4.2.4 Deep active learning

We also test the viability of active learning with DNNs using LSTM networks. An LSTM network architecture for active learning is the same as the baseline LSTM discussed in section 4.1.2, a bidirectional LSTM with 20 hidden nodes and two layers with a dropout probability of 0.2 between layers. Active learning with an LSTM follows the single-pass stream-based active learning algorithm's steps discussed in section 3.3.3, but with a slight modification. A portion of the training samples (e.g., 20%) is reserved for the validation step to halt the training early if the learning stagnates for a certain number of epochs, and two different learning rates are used for initial batch training and incremental updates.



(a) initial LR, update LR $= 1 \times 10^{-5}$

(b) initial LR $= 1 \times 10^{-5}$, update LR $= 9 \times 10^{-5}$

Figure 4.8: Train and validation losses of LSTM-based active learners for arousal classification.

This attempt to apply active learning to an LSTM network nonetheless results in unsuccessful training with models' loss functions failing to converge. Past the initial phase where the training happens rather steadily using a learning rate of 1×10^{-5}, the training quickly destabilizes as the model starts to learn incrementally and fails to acquire any new meaningful information until the training terminates after seeing all samples from the steam. Although it is possible to achieve some incremental learning past the initial training by fine-tuning the second learning rate, the model again soon stops learning when new samples are added. Similar to what we observed while training XGBoost-based active learners, these models were likely incapable of escaping from the local optimum they learned during the initial training and continued to fail to extract new knowledge from queried samples as they are only so many compared to the existing training samples. A continual adjustment of learning rate in the update phase or applying heavier weights to newly queried samples may allow the model to learn new information and correctly incorporate that into its future inferences without forgetting previously acquired information, but that would likely warrant an entirely new approach altogether.

Chapter 5. Discussion: Limitations and Future Works

This work points to that an adaptive sampling approach based on stream-based active learning could significantly reduce data collection down to 60% while reaching a comparable performance to using the full dataset in binary emotion recognition task for arousal and valence. This result, in turn, suggests that *an adaptive ESM can decrease user burden in the process of emotion data collection* to result in improved data quality and the balanced distribution of emotions in the resulting dataset. Our results also provide an insight that guaranteeing an active learning model space for exploration during the early stage of learning can lead to better classification performance, given the imbalanced distribution of training data and the small size of the dataset. Nevertheless, our work is not without limitations and can be further refined in several aspects, including the following:

Stopping criterion for active learning It is apparent from the empirical observation of multiple rounds of training active learners that the training can benefit from a stopping criterion, even when the target coverage is set for a model. This is different from the early stopping used for the training of LSTM models, as it seeks to halt the additional data collection from the data stream if the expected benefit from further querying is below a certain threshold. This idea was similarly explored with Conditional Mutual Information (CMI) in a work that applied active learning to activity recognition [2].

Learnable query function The query function parameters α, β, and γ were set manually during our experiments, but not only the model performance but the efficiency of data collection could be improved if the parameters could be learned through iterative update similar to how the model's internal parameters are updated. Our current implementation of active learning does not support the automated tuning of query parameters as learnable values, but future works should explore this possibility.

Limited ecological validity The ecological validity of the experimental findings is limited as we performed the hold-out evaluation of active learners. While our findings are still applicable assuming a scenario involving a centralized inference module that accumulates data from multiple users to determine which instances to query, personalized emotion recognition models and leave-one-subject-out (LOSO) evaluation would be necessary to apply adaptive sampling in a more realistic scenario involving multiple users and edge devices.

The formulation of emotion recognition as a binary classification problem also limits this work's ecological validity as emotions in the wild cannot be captured in mere two-dimensions as noted in previous work on the theory of emotions (see section 2.1). The division of emotions into two categories of low vs. high was employed to enable streamlined active learning experiments to empirically observe its validity for sampling emotions in realistic scenarios, together with more interpretable results. Nonetheless, this artificial divide limits the observation of natural variances present in the original 5-scale measurement of emotions, which is likely still too limited to observe emotions in their most natural forms.

Therefore, future works may utilize more fine-grained categories of emotions for the higher ecological validity of a model, but possibly altogether invent an embedding space of emotions, similar to the *Word2Vec* technique for embedding natural languages to a multi-dimensional vector space [92]. Such an embedding technique could map emotional states and expressions captured with unimodal or multimodal affective data to space where the contexts those emotions occurred are archived linguistically or in any other means of observing and documenting emotions externally.

Need for ESM specific dataset While the adaptive sampling method explored in this work assumes data collection via ESM in studies lasting several days to possibly months, the K-EmoCon dataset is collected in the 10-minute debate scenario, with participants annotating their emotions during a debate in one sitting. Such artificially of the data collection setup for K-EmoCon limits the extensibility of the dataset and the results of this work to realistic ESM scenarios. Given that, validating whether the observed results of this work hold in experiments with other datasets of emotion data collected in the context of longitudinal ESM would be necessary.

Potential signal aliasing While BVP and ECG signals were resampled to 4Hz from 64Hz and 1Hz respectively, this choice of 4Hz as a resampling frequency is a heuristic one, and does not imply the sampling frequency of the true signal, following the *Nyquist–Shannon sampling theorem*, should be 2Hz. Conversely, a substantially higher frequency is likely to capture affective information in physiological signals fully, but such high-frequency data collection is currently unavailable for commercial-grade mobile wearable sensing devices. However, as a workaround, we could utilize a data fusion strategy or an architecture that omits the need to resample to instead use an educated process to combine data from multiple modalities [25].

Refined deep active learning Although our attempt to actively train a deep neural network was unsuccessful, deep active learning is an active area of research, and adopting a more sophisticated methodology to train a deep network via active learning might succeed as other researchers have demonstrated it [122]. For example, active one-shot learning combining ideas from few-shot learning and reinforcement learning was used to recognize handwritten characters [148], and this methodology could be similarly applied for the emotion recognition task. Investigating the application of meta-learning techniques is also a reasonable option assuming the small size of the dataset [44, 98]. We could also consider different query strategies for active learning, such as the query by committee (QBC), if we assume that the active learning occurs in the multi-user & multi-device environment with one active learner per user. Finally, modifying the loss function directly for active learning [49, 151, 155] and the base network architecture [55, 150] are viable options, given our current choice of binary cross-entropy and a simple LSTM network may have been insufficient for the learning of meaningful representation from physiological signals.

Chapter 6. Conclusion

While ESM is widely employed for research on emotions in the wild, randomly triggered ESM questionnaires can cause interruptions and increase the user burden leading to attrition. The inherently imbalanced distribution of emotions in the wild also affects the data quality and emotion recognition models' performance trained with such data. This work investigated the potential of using active learning-based adaptive sampling methods to collect emotions in a hypothetical data collection scenario simulated with the K-EmoCon dataset.

We implemented the stream-based single-pass active learning algorithm and trained XGBoost and LSTM-based emotion recognition models for binary classification of arousal and valence with a 60-second segment of physiological signals, including BVP, ECG, EDA, and HST. We conducted experiments to test how different configurations of an active learning algorithm affect the learning process and the final model performance with the parameterized query function, which enables controlling the model's selectivity, minimum coverage, and weights to sampling strategies.

The results showed that active learning could reduce data collection down to 60% without significantly sacrificing the model performance, particularly if a base model is rebuilt after each training set update. We also empirically observed allowing exploration during the early training of an active learner can benefit its incremental learning, while careless exploration can also be harmful if the data distribution is severely imbalanced. Although we were not successful in our deep active learning with an LSTM network, findings from few-shot learning and reinforcement learning research should be further explored. Finally, future works should consider validating the findings of this work using a dataset constructed with a longitudinal scenario of collecting emotions in the wild with ESM and employ an evaluation strategy suitable for the realistic scenario involving multiple users and devices to allow the generalization of the adaptive sampling method for naturalistic emotions to real-world applications.

Bibliography

[1] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "Decaf: Meg-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.

[2] R. Adaimi and E. Thomaz, "Leveraging active learning and conditional mutual information to minimize data annotation in human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–23, 2019.

[3] P. Adams, M. Rabbi, T. Rahman, M. Matthews, A. Voida, G. Gay, T. Choudhury, and S. Voida, "Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild," in *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, 2014, pp. 72–79.

[4] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 579–586.

[5] A. Althubaiti, "Information bias in health research: Definition, pitfalls, and adjustment methods," *Journal of multidisciplinary healthcare*, vol. 9, p. 211, 2016.

[6] L. B. Baker, "Physiology of sweat gland function: The roles of sweating and sweat composition in human health," *Temperature*, vol. 6, no. 3, pp. 211–259, 2019.

[7] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[8] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[9] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of neuroscience methods*, vol. 190, no. 1, pp. 80–91, 2010.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[11] C. F. Bond and M. Robinson, "The evolution of deception," *Journal of nonverbal behavior*, vol. 12, no. 4, pp. 295–307, 1988.

[12] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[13] S. Brave and C. Nass, "Emotion in human-computer interaction," *Human-computer interaction fundamentals*, vol. 20094635, pp. 53–68, 2009.

[14] C. Breazeal, "Emotion and sociable humanoid robots," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 119–155, 2003.

[15] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[16] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 205–211.

[17] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[18] J. T. Cacioppo, G. G. Berntson, J. T. Larsen, K. M. Poehlmann, T. A. Ito, *et al.*, "The psychophysiology of emotion," *Handbook of emotions*, vol. 2, pp. 173–191, 2000.

[19] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive behavioural systems*, Springer, 2012, pp. 144–157.

[20] A. Caputo, "Social desirability bias in self-reported well-being measures: Evidence from an online survey," *Universitas Psychologica*, vol. 16, no. 2, pp. 245–255, 2017.

[21] J. M. Carroll and J. A. Russell, "Do facial expressions signal specific emotions? judging emotion from the face in context.," *Journal of personality and social psychology*, vol. 70, no. 2, p. 205, 1996.

[22] R. T. Cauldwell, "Where did the anger go? the role of context in interpreting emotion in speech," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[23] E. Cerin, A. Szabo, and C. Williams, "Is the experience sampling method (esm) appropriate for studying pre-competitive emotions?" *Psychology of Sport and Exercise*, vol. 2, no. 1, pp. 27–45, 2001.

[24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[25] J.-H. Choi and J.-S. Lee, "Embracenet: A robust deep learning architecture for multimodal classification," *Information Fusion*, vol. 51, pp. 259–270, 2019.

[26] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. ARTICLE, pp. 2493–2537, 2011.

[27] D. Colombo, C. Suso-Ribera, J. Fernández-Álvarez, P. Cipresso, A. Garcia-Palacios, G. Riva, and C. Botella, "Affect recall bias: Being resilient by distorting reality," *Cognitive Therapy and Research*, pp. 1–13, 2020.

[28] L. Constantine and H. Hajj, "A survey of ground-truth in emotion data annotation," in *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, IEEE, 2012, pp. 697–702.

[29] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures.," *Emotion*, vol. 18, no. 1, p. 75, 2018.

[30] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, 2018. DOI: https://doi.org/10.1109/TAFFC.2018.2884461.

[31] A. Cowen, D. Sauter, J. L. Tracy, and D. Keltner, “Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression,” *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 69–90, 2019.

[32] A. S. Cowen and D. Keltner, “Self-report captures 27 distinct categories of emotion bridged by continuous gradients,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, E7900–E7909, 2017.

[33] M. Csikszentmihalyi and R. Larson, “Validity and reliability of the experience-sampling method,” in *Flow and the foundations of positive psychology*, Springer, 2014, pp. 35–54.

[34] R. J. Davidson, “On emotion, mood, and related affective constructs,” *The nature of emotion: Fundamental questions*, pp. 51–55, 1994.

[35] H.-J. De Vuyst, E. Dejonckheere, K. Van der Gucht, and P. Kuppens, “Does repeatedly reporting positive or negative emotions in daily life have an impact on the level of emotional experiences and depressive symptoms over time?” *PloS one*, vol. 14, no. 6, e0219121, 2019.

[36] S. Du, Y. Tao, and A. M. Martinez, “Compound facial expressions of emotion,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, E1454–E1462, 2014.

[37] B. Egloff, A. Tausch, C.-W. Kohlmann, and H. W. Krohne, “Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure,” *Motivation and emotion*, vol. 19, no. 2, pp. 99–110, 1995.

[38] G. Eisele, H. Vachon, G. Lafit, P. Kuppens, M. Houben, I. Myin-Germeys, and W. Viechtbauer, “The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population,” *Assessment*, p. 1 073 191 120 957 102, 2020.

[39] P. Ekman, “An argument for basic emotions,” *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[40] ——, “Are there basic emotions?,” 1992.

[41] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[42] P. Ekman and D. Keltner, “Universal facial expressions of emotion,” *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, pp. 27–46, 1997.

[43] T. Fawcett, “Roc graphs: Notes and practical considerations for researchers,” *Machine learning*, vol. 31, no. 1, pp. 1–38, 2004.

[44] C. Finn, A. Rajeswaran, S. Kakade, and S. Levine, “Online meta-learning,” *arXiv preprint arXiv:1902.08438*, 2019.

[45] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang, “Predicting human interruptibility with sensors,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 12, no. 1, pp. 119–146, 2005.

[46] M. G. Frank and E. Svetieva, “Microexpressions and deception,” in *Understanding facial expressions in communication*, Springer, 2015, pp. 227–242.

[47] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[48] N. H. Frijda *et al.*, "Varieties of affect: Emotions and episodes, moods, and sentiments.," 1994.

[49] Y. Geifman and R. El-Yaniv, "Selectivenet: A deep neural network with an integrated reject option," *arXiv preprint arXiv:1901.09192*, 2019.

[50] S. Ghosh, N. Ganguly, B. Mitra, and P. De, "Towards designing an intelligent experience sampling method for emotion detection," in *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, IEEE, 2017, pp. 401–406.

[51] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2008, pp. 1–6.

[52] T. Goetz, M. Bieg, and N. C. Hall, "Assessing academic emotions via the experience sampling method," in *Methodological advances in research on emotion and education*, Springer, 2016, pp. 245–258.

[53] J. J. Gross and O. P. John, "Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being.," *Journal of personality and social psychology*, vol. 85, no. 2, p. 348, 2003.

[54] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[55] G. K. Gudur, P. Sundaramoorthy, and V. Umaashankar, "Activeharnet: Towards on-device deep bayesian active learning for human activity recognition," in *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications*, 2019, pp. 7–12.

[56] J. Healey, "Recording affect in the field: Towards methods and metrics for improving ground truth labels," in *International conference on affective computing and intelligent interaction*, Springer, 2011, pp. 107–116.

[57] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[59] W. Hofmann, R. F. Baumeister, G. Förster, and K. D. Vohs, "Everyday temptations: An experience sampling study of desire, conflict, and self-control.," *Journal of personality and social psychology*, vol. 102, no. 6, p. 1318, 2012.

[60] V. Hollis, A. Konrad, A. Springer, M. Antoun, C. Antoun, R. Martin, and S. Whittaker, "What does all this data mean for my future mood? actionable analytics and targeted reflection for emotional well-being," *Human–Computer Interaction*, vol. 32, no. 5-6, pp. 208–267, 2017.

[61] K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "Cstress: Towards a gold standard for continuous stress assessment in the mobile environment," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 493–504.

[62] Y. Hu, C. Converse, M. Lyons, and W. Hsu, "Neural control of sweat secretion: A review," *British Journal of Dermatology*, vol. 178, no. 6, pp. 1246–1256, 2018.

[63] R. E. Jack, C. Crivelli, and T. Wheatley, "Data-driven methods to diversify knowledge of human psychology," *Trends in cognitive sciences*, vol. 22, no. 1, pp. 1–5, 2018.

[64] A. Kapoor and E. Horvitz, “Experience sampling for building predictive user models: A comparative study,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 657–666.

[65] S. Katsigiannis and N. Ramzan, “Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[66] D. Keltner, “Toward a consensual taxonomy of emotions,” *Cognition and Emotion*, vol. 33, no. 1, pp. 14–19, 2019.

[67] D. Keltner and D. T. Cordaro, “Understanding multimodal emotional expressions: Recent advances in basic emotion theory,” *The science of facial expression*, pp. 57–75, 2017.

[68] D. Keltner and J. Haidt, “Social functions of emotions at four levels of analysis,” *Cognition & Emotion*, vol. 13, no. 5, pp. 505–521, 1999.

[69] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, “Emotional expression: Advances in basic emotion theory,” *Journal of nonverbal behavior*, pp. 1–28, 2019.

[70] D. Keltner, J. L. Tracy, D. Sauter, and A. Cowen, “What basic emotion theory really says for the twenty-first century study of emotion,” *Journal of nonverbal behavior*, vol. 43, no. 2, pp. 195–201, 2019.

[71] Z. D. King, J. Moskowitz, B. Egilmez, S. Zhang, L. Zhang, M. Bass, J. Rogers, R. Ghaffari, L. Wakschlag, and N. Alshurafa, “Micro-stress ema: A passive sensing framework for detecting in-the-wild stress in pregnant mothers,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–22, 2019.

[72] M. Kipp and J.-C. Martin, “Gesture and emotion: Can basic gestural form features discriminate emotions?” In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, IEEE, 2009, pp. 1–8.

[73] G. A. van Kleef, A. Cheshin, A. H. Fischer, and I. K. Schneider, “The social nature of emotions,” *Frontiers in Psychology*, vol. 7, p. 896, 2016.

[74] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis; using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[75] D.-S. Kwon, Y. K. Kwak, J. C. Park, M. J. Chung, E.-S. Jee, K.-S. Park, H.-R. Kim, Y.-M. Kim, J.-C. Park, E. H. Kim, *et al.*, “Emotion interaction system for a service robot,” in *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, IEEE, 2007, pp. 351–356.

[76] R. J. Larsen and E. Diener, “Affect intensity as an individual difference characteristic: A review,” *Journal of Research in personality*, vol. 21, no. 1, pp. 1–39, 1987.

[77] R. Larson and M. Csikszentmihalyi, “The experience sampling method,” in *Flow and the foundations of positive psychology*, Springer, 2014, pp. 21–34.

[78] H. C. Lench and L. J. Levine, “Motivational biases in memory for emotions,” *Cognition and emotion*, vol. 24, no. 3, pp. 401–418, 2010.

[79] L. J. Levine and M. A. Safer, “Sources of bias in memory for emotions,” *Current directions in psychological science*, vol. 11, no. 5, pp. 169–173, 2002.

[80] B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades, “Facing the cold start problem in recommender systems,” *Expert Systems with Applications*, vol. 41, no. 4, pp. 2065–2073, 2014.

[81] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, “Moodscope: Building a mood sensor from smartphone usage patterns,” in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 2013, pp. 389–402.

[82] C. H. Lin, M. Mausam, and D. S. Weld, “Active learning with unbalanced classes and example-generation queries.,” in *HCOMP*, 2018, pp. 98–107.

[83] J. Liono, F. D. Salim, N. van Berkel, V. Kostakos, and A. K. Qin, “Improving experience sampling with multi-view user-driven annotation prediction,” in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*, IEEE, 2019, pp. 1–11.

[84] C. L. Lisetti and F. Nasoz, “Using noninvasive wearable computers to recognize human emotions from physiological signals,” *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 11, p. 929 414, 2004.

[85] V. Markova, T. Ganchev, and K. Kalinkov, “Clas: A database for cognitive load, affect and stress recognition,” in *2019 International Conference on Biomedical Innovations and Applications (BIA)*, IEEE, 2019, pp. 1–4.

[86] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, “Learning deep physiological models of affect,” *IEEE Computational intelligence magazine*, vol. 8, no. 2, pp. 20–33, 2013.

[87] R.-E. Mastoras, D. Iakovakis, S. Hadjidimitriou, V. Charisis, S. Kassie, T. Alsaadi, A. Khandoker, and L. J. Hadjileontiadis, “Touchscreen typing pattern analysis for remote detection of the depressive tendency,” *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[88] J. D. Mayer, D. R. Caruso, and P. Salovey, “Emotional intelligence meets traditional standards for an intelligence,” *Intelligence*, vol. 27, no. 4, pp. 267–298, 1999.

[89] D. McDuff, M. Amr, and R. El Kaliouby, “Am-fed+: An extended dataset of naturalistic facial expressions collected in everyday settings,” *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 7–17, 2018.

[90] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, “The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2011.

[91] A. Mehrotra, F. Tsapeli, R. Hendley, and M. Musolesi, “Mytraces: Investigating correlation and causation between users’ emotional states and mobile phone interaction,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–21, 2017.

[92] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[93] T. Miu, P. Missier, and T. Plötz, “Bootstrapping personalised human activity recognition models using online active learning,” in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, IEEE, 2015, pp. 1138–1147.

[94] A. Mollahosseini, B. Hasani, and M. H. Mahoor, “Affectnet: A database for facial expression, valence, and arousal computing in the wild,” *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[95] M. B. Morshed, K. Saha, R. Li, S. K. D'Mello, M. De Choudhury, G. D. Abowd, and T. Plötz, "Prediction of mood instability with passive sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–21, 2019.

[96] M. T. Motley and C. T. Camden, "Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting," *Western Journal of Communication (includes Communication Reports)*, vol. 52, no. 1, pp. 1–22, 1988.

[97] I. Myin-Germeys, Z. Kasanova, T. Vaessen, H. Vachon, O. Kirtley, W. Viechtbauer, and U. Reininghaus, "Experience sampling methodology in mental health research: New insights and technical developments," *World Psychiatry*, vol. 17, no. 2, pp. 123–132, 2018.

[98] A. Nagabandi, C. Finn, and S. Levine, "Deep online learning via meta-learning: Continual adaptation for model-based rl," *arXiv preprint arXiv:1812.07671*, 2018.

[99] L. Nummenmaa, E. Glerean, R. Hari, and J. K. Hietanen, "Bodily maps of emotions," *Proceedings of the National Academy of Sciences*, vol. 111, no. 2, pp. 646–651, 2014.

[100] J. Ocumpaugh, "Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual," *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences*, vol. 60, 2015.

[101] K.-J. Oh, D. Lee, B. Ko, and H.-J. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation," in *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, IEEE, 2017, pp. 371–375.

[102] J. van Os, S. Verhagen, A. Marsman, F. Peeters, M. Bak, M. Marcelis, M. Drukker, U. Reininghaus, N. Jacobs, T. Lataster, *et al.*, "The experience sampling method as an mhealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice," *Depression and anxiety*, vol. 34, no. 6, pp. 481–493, 2017.

[103] G. Paltoglou and M. Thelwall, "Seeing stars of valence and arousal in blog posts," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 116–123, 2012.

[104] C. Y. Park, *Cheulyop/pyteap: Pyteap v0.1.2*, version v0.1.2, Dec. 2020. DOI: 10.5281/zenodo.4319682. [Online]. Available: https://doi.org/10.5281/zenodo.4319682.

[105] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, p. 293, Sep. 2020. DOI: 10.1038/s41597-020-00630-y.

[106] V. Pejovic and M. Musolesi, "Interruptme: Designing intelligent prompting mechanisms for pervasive applications," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 897–908.

[107] R. W. Picard, *Affective computing*. MIT press, 2000.

[108] ——, "Future affective technology for autism and emotion communication," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3575–3584, 2009.

[109] R. W. Picard and J. Healey, "Affective wearables," *Personal Technologies*, vol. 1, no. 4, pp. 231–240, 1997.

[110] R. W. Picard, E. Vyzas, and J. Healey, “Toward machine emotional intelligence: Analysis of affective physiological state,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[111] M. Pielot, B. Cardoso, K. Katevas, J. Serrà, A. Matic, and N. Oliver, “Beyond interruptibility: Predicting opportune moments to engage mobile phone users,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–25, 2017.

[112] R. Plutchik, “Emotions: A general psychoevolutionary theory,” *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.

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

[114] ——, “Meld: A multimodal multi-party dataset for emotion recognition in conversations,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[115] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[116] B. Reeves and C. I. Nass, *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press, 1996.

[117] S. Rosenthal, A. K. Dey, and M. Veloso, “Using decision-theoretic experience sampling to build personalized mobile phone interruption models,” in *International Conference on Pervasive Computing*, Springer, 2011, pp. 170–187.

[118] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[119] J. A. Russell, “A circumplex model of affect.,” *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[120] M. A. Safer, L. J. Levine, and A. L. Drapalski, “Distortion in memory for emotions: The contributions of personality and post-event knowledge,” *Personality and Social Psychology Bulletin*, vol. 28, no. 11, pp. 1495–1507, 2002.

[121] K. Saha, L. Chan, K. De Barbaro, G. D. Abowd, and M. De Choudhury, “Inferring mood instability on social media by leveraging ecological momentary assessments,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–27, 2017.

[122] D. Sahoo, Q. Pham, J. Lu, and S. C. Hoi, “Online deep learning: Learning deep neural networks on the fly,” *arXiv preprint arXiv:1711.03705*, 2017.

[123] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets,” *PloS one*, vol. 10, no. 3, e0118432, 2015.

[124] P. Salovey and J. D. Mayer, “Emotional intelligence,” *Imagination, cognition and personality*, vol. 9, no. 3, pp. 185–211, 1990.

[125] K. R. Scherer, “What are emotions? and how can they be measured?” *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[126] C. N. Scollon, C.-K. Prieto, and E. Diener, “Experience sampling: Promises and pitfalls, strength and weaknesses,” in *Assessing well-being*, Springer, 2009, pp. 157–180.

[127] D. Sculley, “Online active learning methods for fast label-efficient spam filtering.,” in *CEAS*, vol. 7, 2007, p. 143.

[128] B. Settles, *Active Learning.* Morgan & Claypool Publishers, 2012, ISBN: 1608457257.

[129] S. Shalev-Shwartz *et al.*, “Online learning and online convex optimization,” *Foundations and trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.

[130] J. Shao, Q. Wang, and F. Liu, “Learning to sample: An active learning framework,” in *2019 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2019, pp. 538–547.

[131] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, “A dataset of continuous affect annotations and physiological signals for emotion analysis,” *Scientific Data*, vol. 6, no. 1, pp. 1–13, 2019.

[132] S. Shiffman, A. A. Stone, and M. R. Hufford, “Ecological momentary assessment,” *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[133] M. N. Shiota, B. Campos, C. Oveis, M. J. Hertenstein, E. Simon-Thomas, and D. Keltner, “Beyond happiness: Building a science of discrete positive emotions.,” *American Psychologist*, vol. 72, no. 7, p. 617, 2017.

[134] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2011.

[135] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, “Toolbox for emotional feature extraction from physiological signals (teap),” *Frontiers in ICT*, vol. 4, p. 1, 2017.

[136] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, “Ascertain: Emotion and personality recognition using commercial sensors,” *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.

[137] Y. Suhara, Y. Xu, and A. Pentland, “Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks,” in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 715–724.

[138] A. Thieme, D. Belgrave, and G. Doherty, “Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems,” *ACM Trans. Comput. Hum. Interact.*, vol. 27, 2020.

[139] S. B. Thrun, “Efficient exploration in reinforcement learning,” 1992.

[140] K. P. Truong, D. A. Van Leeuwen, and F. M. De Jong, “Speech-based recognition of self-reported and observed emotion in a dimensional space,” *Speech communication*, vol. 54, no. 9, pp. 1049–1063, 2012.

[141] H. Vachon, M. Bourbousson, T. Deschamps, J. Doron, S. Bulteau, A. Sauvaget, and V. Thomas-Ollivier, “Repeated self-evaluations may involve familiarization: An exploratory study related to ecological momentary assessment designs in patients with major depressive disorder,” *Psychiatry Research*, vol. 245, pp. 99–104, 2016.

[142] G. A. Van Kleef, *The interpersonal dynamics of emotion.* Cambridge University Press, 2016.

[143] S. J. Verhagen, L. Hasmi, M. Drukker, J. van Os, and P. A. Delespaul, "Use of the experience sampling method in the context of clinical trials," *Evidence-based mental health*, vol. 19, no. 3, pp. 86–89, 2016.

[144] F. Villaro-Dixon, T. Pun, and G. Chanel, *Gijom/teap: Teap v0.1 (alpha)*, version v0.1-alpha, Sep. 2016. DOI: 10.5281/zenodo.154147. [Online]. Available: https://doi.org/10.5281/zenodo.154147.

[145] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, 2014, pp. 3–14.

[146] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing twitter" big data" for automatic emotion identification," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, IEEE, 2012, pp. 587–592.

[147] R. L. Widdershoven, M. Wichers, P. Kuppens, J. A. Hartmann, C. Menne-Lothmann, C. J. Simons, and J. A. Bastiaansen, "Effect of self-monitoring through experience sampling on emotion differentiation in depression," *Journal of affective disorders*, vol. 244, pp. 71–77, 2019.

[148] M. Woodward and C. Finn, "Active one-shot learning," *arXiv preprint arXiv:1702.06559*, 2017.

[149] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2017, pp. 248–255.

[150] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 351–360.

[151] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 93–102.

[152] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, Springer, 2014, pp. 818–833.

[153] J. M. Zelenski and R. J. Larsen, "The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data," *Journal of Research in Personality*, vol. 34, no. 2, pp. 178–197, 2000.

[154] B. Zhang, G. Essl, and E. Mower Provost, "Automatic recognition of self-reported and perceived emotion: Does joint modeling help?" In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 217–224.

[155] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on knowledge and data engineering*, vol. 18, no. 1, pp. 63–77, 2005.

Acknowledgment

Thanks to Prof. Uichin Lee for all guidance he had given me over the last two and a half years in life and academics alike. Thanks to my family for their unconditional love and understanding, allowing me to endeavor in any hardship. Thanks to friends and colleagues for being companions on my side, kind and supportive despite my shortcomings. Finally, thanks to all who contributed their valuable time and data in constructing the K-EmoCon dataset. Without you all, this would not have been possible.