Modeling User Affect Using Interaction Events

by

Areej Alhothali

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2011

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Emotions play a significant role in many human mental activities, including decision-making, motivation, and cognition. Various intelligent and expert systems can be empowered with emotionally intelligent capabilities, especially systems that interact with humans and mimic human behaviour. However, most current methods in affect recognition studies use intrusive, lab-based, and expensive tools which are unsuitable for real-world situations. Inspired by studies on keystrokes dynamics, this thesis investigates the effectiveness of diagnosing users' affect through their typing behaviour in an educational context. To collect users' typing patterns, a field study was conducted in which subjects used a dialogue-based tutoring system built by the researcher. Eighteen dialogue features associated with subjective and objective ratings for users' emotions were collected. Several classification techniques were assessed in diagnosing users' affect, including discrimination analysis, Bayesian analysis, decision trees, and neural networks. An artificial neural network approach was ultimately chosen as it yielded the highest accuracy compared with the other methods. To lower the error rate, a hierarchical classification was implemented to first classify user emotions based on their valence (positive or negative) and then perform a finer classification step to determining which emotions the user experienced (delighted, neutral, confused, bored, and frustrated). The hierarchical classifier was successfully able to diagnose users' emotional valence, while it was moderately able to classify users' emotional states. The overall accuracy obtained from the hierarchical classifier significantly outperformed previous dialogue-based approaches and in line with some affective computing methods.

Acknowledgements

I would like to thank everyone who made this work possible starting by my supervisor Prof.Chrysanne DiMarco for her guidance and encouragement. I would also to thank the participants for taking time to take part in this study. I would also like to thank Tracy Xiong and Vivian Wing-Sheung Chan for rating the participants' emotions. Finally, I would like to thank Prof.Edward Lank and Prof.Peter Van Beek for being my thesis readers.

Dedication

This thesis is dedicated to my parents, grandmother, sisters and brothers who have supported me all the way since the beginning of my studies.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Understanding of user emotions would be a very advantageous capability for many types of intelligent computer systems. This is especially true for on-line educational systems, health educational systems, and persuasive technology systems, which all rely on users' emotions. Several studies in various disciplines, including cognitive science, and psychology, have shown that emotions influence individuals' mental activities such as cognition, perception, and creativity [29].

Among educators, it is generally believed that students who experience negative emotions tend to be less interested, and less able to learn new educational information, whereas students who are in positive affective states tend to be more receptive to learning new information [22][23]. Experienced tutors, teachers, and academics are naturally aware of the correlation between cognition and emotion. Thus, they are accustomed to adapting their teaching style based on their students' emotions. [23]

In the affective computing field, which focuses on building emotionally intelligent computer systems able to recognize and respond to users' emotions, researchers typically diagnose users' emotions by analyzing their physiological or behavioural patterns. Physiological approaches diagnose a user's affective states by measuring physiological responses, such as skin conductance, heart rate, blood pressure, pupillary response, respiration rate, breathing rate, or brain-wave patterns. In contrast, behavioural approaches map a user's emotions to his/her physical (behavioural) reactions such as facial expressions, body gesture, body pressure, and voice intonation.

Despite the accuracy of physiological and behavioural methods, both approaches rely on intrusive, and expensive lab-based tools that are not common as peripherals in personal computers. Thus, applying and using these tools are not feasible for everyday use. Also, the

impact of being monitored by wearable devices or video cameras may influence users' behavioural or physiological reactions, which makes it difficult to attribute physiological and behavioural changes to users' emotions [1]. Moreover, diagnosing users' emotions through behavioural patterns relies on apparent or exaggerated reactions that do not represent the typical interactivity that users experience while engaging with regular computer systems.

Therefore, it is advantageous to investigate the effectiveness of using non-obtrusive and easy-to-use methods to diagnose users' emotions. In the field of Affective Tutoring Systems (ATS), several studies have modeled users' affect in intelligent tutoring system (ITS). However, most of these studies have used physiological or behavioural approaches. This thesis investigates the effectiveness of using keystroke features when assessing users' emotions during their ongoing interaction with ITS. Unlike related methods in the field, our method is non-intrusive and personalized. As well, it does not require the user to have any additional tools, nor does it require users to perform other tasks.

In the last decade, many studies in the computer security field utilized keystroke features to detect unauthenticated users during the authentication process, by comparing a given typing pattern with the typing pattern of authenticated users. However, less attention has been paid in the affective computing field to investigating the possibility of using typing patterns as an indicator of emotional change. In this study, we define a new set of features that are commonly used in keystroke dynamics and stress detection studies including timing, typing, and response features to classify users' emotions in educational dialogue contexts.

The timing features we used are: session duration and pause rate. The typing features are: typing speed, punctuation rate, capitalization rate, unrelated key rate, key duration, key latency, and deletion rate. The response features are: quality, length, and correctness of the users' answer. In this research, we concerned with emotions that have been shown to be related to the learning process [23][17][29]: delight, neutral, confusion, boredom, and frustration.

To build an emotionally intelligent classifier, the classifier needs to be trained on feature vectors, each associated with an emotion category. To collect the interaction features, an intelligent tutoring system (CompTutor) was built which teaches students various computer-related topics through a written dialogue. CompTutor consists of: a user model, domain model, feedback generator, and diagnosis module. The system automatically computes and extracts the typing features listed above.

A field study was conducted to gather the interaction data of 20 students from the University of Waterloo. The study consisted of two conditions: baseline condition and experimental condition. In the baseline condition, the participants were asked to type part

of a paragraph during a state of no cognitive stress (neutral). In the experimental condition, each participant was asked to interact with CompTutor for 30 minutes by answering the system questions. During both conditions, the participants' typing features were tracked and their behavioural reactions were video-recorded.

After each interaction, the participants were asked to self-report their emotions by answering a short survey to determine their emotions during their ongoing interaction with the system. In addition, two trained judges from the psychology department watched the participants through recorded video and evaluated their emotions according to their behavioural responses. The user-labelled dataset consisted of 544 feature vectors and the judge-labelled dataset consisted of 581 feature vectors. Each vector consisted of 18 features associated with an emotion category.

Six classification methods were assessed in diagnosing users' emotions and emotional valence including: linear discrimination analysis (LDA), quadratic discrimination analysis (QDA), decision trees, Naive Bayes, k-nearest neighbour, and artificial neural networks. The classification accuracy of using the neural network was significantly higher than the other classification methods with an accuracy of 82.82% on the user-labelled dataset, 72.02% on judge1 dataset, and 77.2% on judge2 datasets. However, the accuracy of detecting emotional states was 53.59%, 45.6%, and 53.89% for the user-labelled, judge1, and judge2 dataset, accordingly.

The accuracy of classifying user emotional valence (positive or negative) was significantly higher than distinguishing between the five emotions. Thus, a hierarchical classification was implemented to first classify the features based on users' emotional valence, and secondly to reclassify the data into one of the five emotion categories. The hierarchical classifier outperformed the standard classification techniques as it yielded an average accuracy of 59.37%, 49.74%, and 56.48% for the user-labelled, judge1, and judge2 datasets, respectively.

Using interaction features solely was found to be sufficient to successfully diagnose users' emotional valence (positive, negative) and moderately able to determine emotional states. In comparison with affective computing methods, the overall accuracy obtained from our method outperformed previous dialogue-based methods [9], and is in line with other affective computing methods such as audio-based. Vision-based and multi-modal methods provide higher accuracy, but require additional tools and more computational expense in comparison with solely using interaction features [4].

Chapter 2

Related Research

This chapter consists of two sections: the first section presents the terminology that has been used in this study and gives a brief background about the relationship between emotions and cognition. The second section presents the background work of keystroke dynamics relevant to Computer Security and Affective Computing.

2.1 Emotions Theoretical Background

2.1.1 Terminology

In defining emotions, emotion theorists traditionally follow one of two approaches: the cognitive approach or the behavioural approach. Advocates of the cognitive approach believe that emotions are cognitive responses experienced in the brain, independent of bodily sensations. The behavioural (or physical) approach focuses on physiological responses (e.g., heart rate, blood pressure, and respiration rate) that occur prior to or during an emotional episode [29][32]. Currently, most researchers consider emotions to be a combination of both cognitive and physical responses, where both cognitive thoughts and body chemistry can influence individuals' emotions [29].

In this research the terms *mood*, *affect*, and *emotion* will be used interchangeably. Psychologist and cognitive scientist often use the term affect to refer to emotion or mood. Emotion and mood have are both affective states, but they have different characteristics. Emotion is usually associated with specific causes, causes immediate reactions and and lasts for short periods of time. On the other hand, mood tends to be more subtle, less

intensive, longer lasting and non-specific (either negative or positive). As well, mood does not necessarily lead to physical reaction [29].

2.1.2 Emotion Categorization

Some researchers describe emotions in a descriptive form such as the basic emotions that are common among all human beings: fear, anger, sadness, and joy [10]. Other researchers include more emotions. Plutchik [31] for example differentiated eight emotions: acceptance, anger, anticipation, disgust, joy, fear, sadness, and surprise, while Ekman [10] focused on emotions that can be clearly distinguished through facial expressions: anger, disgust, fear, joy, sadness, and surprise.

Other researchers categorize emotions according to two or more dimensions. For example, Russell's [32] dimensional model is one of the most acknowledged models that categorizes emotions according to two independent dimensions, *valence* and *arousal*, Figure 2.1. The term valence refers to the general description of an emotion: positive/negative or pleasant/un-pleasant. The term arousal refers to a human's momentary level of excitation to a stimulus which is usually described as high arousal or low arousal.

2.1.3 Emotions and Learning

Human's emotions play a key role in most mental activities. Studies in different disciplines such as psychology, cognitive science, and computing, indicate that emotions have an influence on various human mental processes including perceptions, decision-making, creativity, memory, and motivation [6][23][29]. Some studies also show that students who are in a positive emotional state tend to be more highly motivated to learn, pay more attention to the instructor, and retain educational information more easily than those who are in a negative emotional state [19][29]. The correlation between emotions and cognition is well-known to academics, tutors, researchers, and teachers. Although experienced human tutors tend to adapt and tailor their teaching methods according to students' emotions, few on-line educational systems consider users' affect and adapt the teaching style accordingly [23].

Various models have been proposed to identify the basic emotions that students can experience during their learning process. Kort et al. [17] define 30 emotions related to learning which are grouped into five emotion axes with respective degrees of positivity and negativity: anxiety-confidence, boredom-fascination, frustration-euphoria, dispirited-encourage, and terror-enchantment. He also proposed a scientific model relating emotions

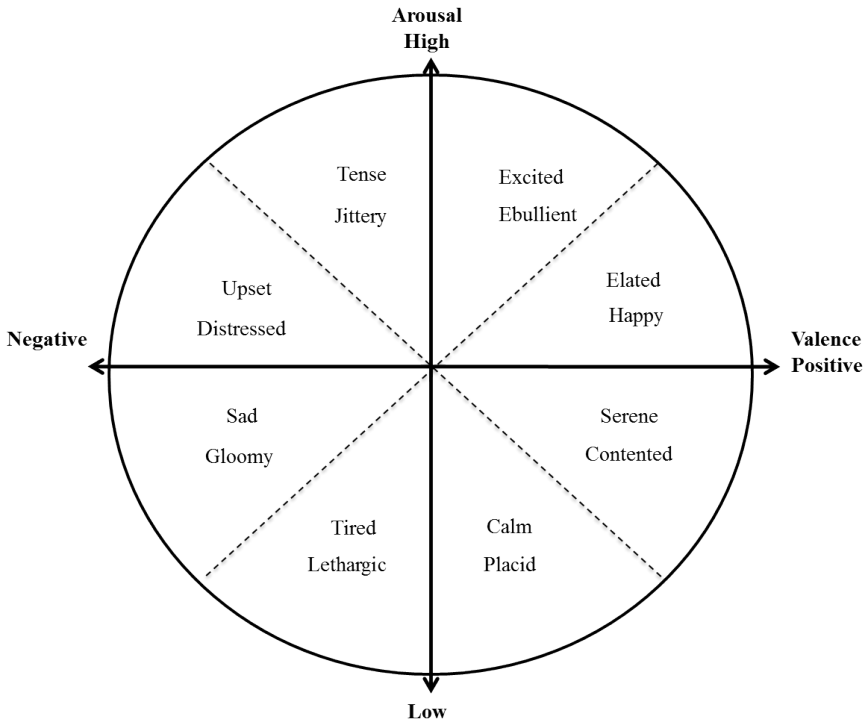

Figure 2.1: Core affect dimensional model of emotions

to cognition and learning. As shown in Figure 2.2, the model of a learning cycle consists of two axes: learning and emotion. The emotion axis consists of two valences: positive emotions (such as satisfaction and curiosity) on the right, and negative emotions (such as frustration and confusion) on the left. The vertical axis represents the constructive learning valence (top), and the un-learning valence (bottom).

Kort et al.'s[17] model suggests that students usually start from the first two quadrants and move in the model contra-clockwise. For example, if students are solving mathematical problem, they will start from the first or second quadrant, either curious or confused about the problem. If they were not able to provide the correct answer, they will move to the third quadrant. But, if they were able to solve the problem, they will move again to the first quadrant. Kort et al.'s research results indicate that learning and emotions are not steady throughout the learning process. However, students usually experience different emotions as they move from un-learning state to constructive learning state.

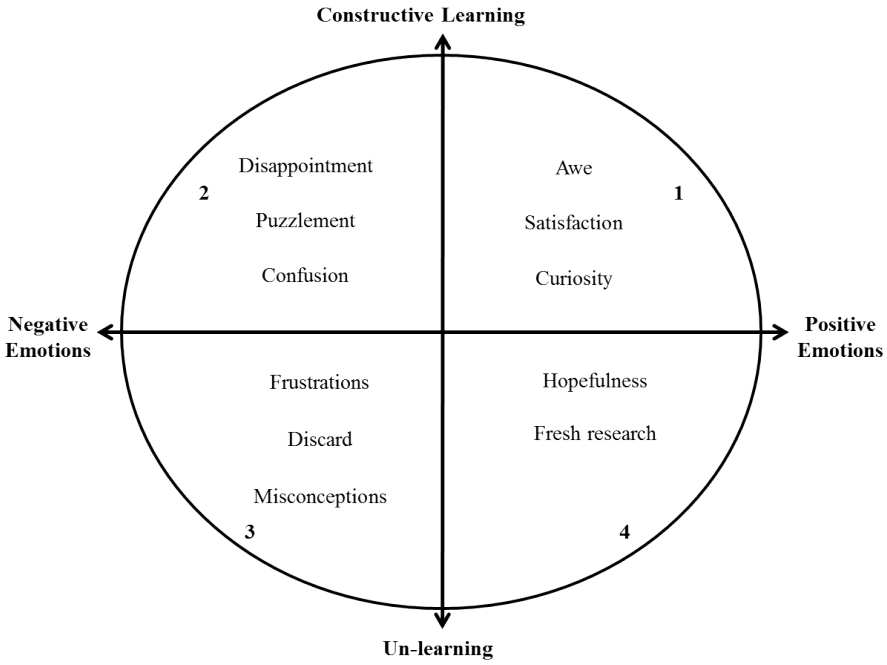

Figure 2.2: Kort's model relating learning to emotions

2.1.4 Affective Computing and Intelligent Tutoring Systems (ITS)

Affective computing focuses on giving computer systems the ability to recognize an individual's emotions and to respond intelligently to the user's emotions [29]. Detecting the user's affective state can provide important information which could be used to improve the usability and the functionality of many applications. For example, if a computer system is able to detect when the user is confused or frustrated, it can guide the user to suitable help, thereby saving time and effort and increasing user satisfaction. Having the capability to diagnose users' emotions would especially benefit expert systems such as tutoring systems, health educational systems, and persuasive technology systems that interact with humans and mimic human reactions.

Recently, affect recognition has become an important aspect of intelligent tutoring systems. In 1997, Rosalind Picard first introduced the term affective tutoring system (ATS) in her book *Affective Computing* [29]. ATS is a term that refers to an intelligent tutoring system that is able to diagnose and tailor the teaching style according to users' affective state. For example, if a user was frustrated while he/she was solving a mathematical problem, the system would provide further explanations or hints. On the other hand, the

system could assign more challenging tasks if a user was bored during a tutoring session.

2.1.5 Affect Measurement Methods

Many methods have been investigated to diagnose emotions which are classified into two categories: physiological and behavioural approaches. The correlation between physiological signals and emotions was first addressed by Ekman et al. [11], who hypothesized that physiological signals can be used as evidence for emotional change. They defined several physiological characteristics as patterns for specific emotions. Subsequently, several researchers in the field of Affective Computing studied the feasibility of using physiological signals such as body temperature, heart rate, blood pressure, pupillary response, respiration rate, or brain-wave patterns to diagnose emotions.

On the other hand, advocates of behavioural approaches believe that humans often recognize each other's emotions by using their own empirical knowledge to evaluate faces, voices, and body signals. Similarly, the majority of affect recognition studies exploit the users' behavioural patterns, including facial expression, body gesture, and voice intonation to determine their emotional states. Yet, face and vocal analysis are the most prominent methods studied in the field of Affective Computing.

Most of the proposed methods use individuals' physiological or behavioural patterns to diagnose their emotions use intrusive and expensive lab-based tools that are not commonly used as peripheral devices of personal computers. Using these tools therefore is not feasible in everyday life. Also, some physiological responses that are known to be associated with specific emotions may arise for different reasons without corresponding to any emotions [29]. For example, having high blood pressure is a physiological sign that is associated with experiencing negative emotions. However, people who are doing some physical activities might have higher blood pressure and not be experiencing any negative emotions.

In addition, several studies in the medical field indicate that monitoring individuals using wearable devices may influence their physiological responses. For example, Ayman et al.[1] indicate that 20-25% of their patients suffer from a syndrome called "white coat hypertension" where they experience fear of being in a physician's office or clinical setting which causes acute high blood pressure. Similarly, being monitored in a laboratory setting using wearable devices or video cameras may influence the participants' physiological or behavioural responses which affect the reliability of the data.

Moreover, diagnosing users' emotions using their behavioural patterns relies on obvious or exaggerated physical reactions which do not represent the typical interactivity that users usually experience when they interact with ordinary computer systems. For example,

people usually don't make clear facial expressions such as smiling and scowling when they are using on-line educational systems as when they are watching movies or chatting with friends. Also, it is difficult to attribute some physical responses to specific emotions since different people have different ways of expressing their feelings, e.g., while some people tend to lean backward when they are nervous, others might tend forward.

For the above-mentioned reasons, there is a need to develop new effective, easy-to-use, accessible, and non-intrusive tools to diagnose individuals' emotions. Inspired by keystroke dynamics studies, we investigated the feasibility of using typing features to diagnose emotional states. Our proposed method fulfills our arguments of being personalized, non-intrusive, and easy-to-use. As well, this method exploits the interaction data to diagnose user's emotions without requiring the user to do any additional tasks.

2.2 Related Research

2.2.1 Keystroke Dynamics and Authentication Systems

The majority of studies in keystroke dynamics are for authentication and verification purposes. Since Gaines et al. [12] first proposed an approach using keystroke dynamics to verify users' identity, typing patterns (or keystroke-dynamics) have been studied extensively for security applications, to enhance the authentication process by comparing the supplied typing pattern with a previously constructed typing pattern. Typically, researchers in keystroke dynamics studies have built a reference model from either the users' spontaneously generated text, or predefined documents. Then, using machine learning techniques the classifier compares the current typing pattern with the reference model. This procedure is usually implemented during the authentication process (when the user entered his/her password), or after the authentication process (during the ongoing interaction).

Monorose and Rubin [25] built an automated classifier that used keystroke features including keystroke timing and key latency to detect unsuitability in users' typing patterns to enhance the authentication process. The typing features were extracted from both predefined text (fixed text) and spontaneously generated text (free text). Their proposed method yielded a 48.9% accuracy-recognition rate for a population of 31 users. Monorose et al. [26], in another study, suggested that individuals' typing patterns are not stable, and change according to their environment, stress level, and cognitive function.

A recent study by Villani et al. [37] utilized the users' interaction data, including mouse activity, keystroke dynamics and timing characteristics to generate a reference model. The

analyzed data consisted of both free text and predefined text. The extracted features included non-letter keys rate, mouse activity rate, pause rate, and input rate. After data collection, a classification based on Euclidean distance was used to compare both current patterns with the previously constructed model. The resulting evaluation of the model showed that when the same type of keyboard was used for both models, the correct recognition rate ranged from 98.3% to 99.5% for a population of 36 users. However, it dropped to 59% when participants used different types of keyboards.

2.2.2 Affective Computing and Keystroke Dynamics

Recently Vizer, Zahou and Sears [38] proposed a novel approach to diagnose individuals' cognitive and physical functions using keystroke dynamics. Unlike the related approaches used to measure cognitive and physical stress, this approach used unobtrusive tools that were adequate for continuous tracking. Measuring cognitive and physical functionality usually takes place in a clinical setting using obtrusive tools not designed for continuous monitoring. Their approach utilized the users' everyday interactions to detect changes in their cognitive and physical functions.

The experiment consisted of control (no stress) condition, cognitive stress condition, and physical stress condition where the participants were asked to provide a text sample under each condition. The extracted features included timing, keystroke (such as key rate, deletion rate and pause rate), and linguistic features. The collected data was analyzed using several machine learning techniques: decision trees, support vector machine, k-nearest neighbour, AdaBoost, and artificial neural networks.

The results of the study indicated that there were recognizable changes in users' typing patterns under cognitive and physical stress.The classification accuracy is 75% in the case of cognitive stress, which was close to the accuracy level provided by affective computing methods. However, classification accuracy for physical stress is 62.5%. The results of the study suggest that individual's typing patterns provide valuable information to detect the presence of cognitive and physical stress. This result in turn has motivated affective computing researchers to investigate the efficiency of using typing patterns to diagnose emotions.

Zimmermann et al. [44] described a field study where they evaluated the use of keyboard and mouse interactions to detect users' emotions. This study used the categorical labelling scheme that uses emotional valence and arousal dimensions resulting in four different mood states, namely PVHA (positive valence and high arousal), PVLA (positive valence low arousal), NVHA (negative valence high arousal), and NVLA (negative valence low arousal).

In the experiment, participants were asked to shop on-line for office supplies while they were experiencing various affective states that were induced using video clips.

Participants' emotions were assessed using physiological sensors that measured their respiration rate, pulse rate, and skin conductance. They were also asked to self-report their emotional states by using the Self-Assessment-Manikin (SAM) [2], which consisted of graphical manikin that each represents score in the valance and arousal dimension (see Figure 2.1). Zimmermann et al's classifier was able to distinguish between neutral and other emotional states, but was not able to distinguish between the other four induced states.

In a similar study, Tsihrintzis et al. [36] investigated the possibility of improving the accuracy of visual-facial emotion recognition using keystroke information. They conducted two studies to evaluate both vision-based modalities and multi-modalities that use facial features and keystroke information to detect users' emotions. The study focused on Ekman's six basic emotions: happiness, sadness, surprise, anger, disgust, and neutral. Participants were shown a set of pictures indicating various emotions, and each participant was asked to classify the emotion indicated by each picture. The result of the first experiment suggested that facial expressions can provide important evidence to diagnose some emotions, such as happiness, neutral and surprise. However, the accuracy of using facial expressions solely to diagnose anger and sadness was low compared to other emotions.

In the second experiment, both keystroke features and facial features were used to assess individuals' emotions. The keystroke features that were investigated in this research were: typing speed rate, pause rate, and deletion rate. The results of the experiments showed that by using keystroke information solely participants were able to detect anger with accuracy of 74%, sadness with 57% accuracy, and neutral with 65% accuracy. However, participants were not able to recognize surprise and disgust using keystroke information. The result of the survey showed that most of the participants tended to make more mistakes and use the backspace key more frequently when they were experiencing negative feelings (nervous, or angry). They tended to type faster when they were experiencing positive feelings and they typed slower when they were experiencing negative feelings.

2.2.3 Tutoring Systems and Typing Features

In ATS research field, investigator have built models of affect for multiple emotions, single emotion, and overlapping emotions. The next section presents the state-of-the-art of modelling user affect in intelligent-tutoring systems using keystroke features.

2.2.3.1 Multi-Emotions Recognition

D'Mello et al. [3] [9] evaluated the effectiveness of using conversational cues to detect users' current state of emotion during interaction with Auto-Tutor (a dialogue-based intelligent tutoring system). The interaction data mined from Auto-Tutor's log file included temporal information, response information, answer quality, tutor directness, and tutor feedback. The researchers in this study conducted an experiment to map the dialogue features to users' emotions. The study focused on six pre-defined emotions that directly influence the learning process. These included: boredom, confusion, delight, flow, frustration, and surprise. The tutoring sessions were video-recorded and evaluated by four judges: the user, a peer, and two trained judges. The researchers evaluated different machine-learning techniques to diagnose users' emotions, including simple logistic regression, decision trees, and Bayesian classification.

The results of the study showed that the reliability of using dialogue features to automatically diagnose users' emotions was overall significantly lower than the trained judges, and slightly lower than the novice judges. Moreover, the reliability of using dialogue features was significantly lower than novice judges in detecting moments of delight and surprise. However, using multiple regression analysis the researchers in this study found a significant correlation between certain typing features and emotions. On the basis of these results, the authors concluded that the accuracy of detecting emotions using keystroke features is significantly lower than affective computing behavioural and physiological methods.

This research yielded many valuable results in affective computing. The researchers studied all the possible dialogue features that might help to diagnose users' emotions. Surprisingly though, they failed to consider some additional typing features that proved to provide significant information to predict users' emotional states such as typing speed, pause rate, deletion rate, and use of unrelated key rate [36] [42].

In D'Mello et al.'s study, the reliability of the self-labelling procedure is questionable as the participants were asked to identify their emotions after the experiment when they were watching their recorded sessions, which might have some impact on the validity of their labelling, as the students might have forgotten what emotions they experienced during the tutoring sessions.

2.2.3.2 Single-Emotion Recognition

The primary focus of effect detection studies is on building an intelligent system that differentiates between several emotions, but some studies focused on detecting a single

emotion. Muldner et al. [27] studied the effectiveness of using dialogue features and pupillary response in detecting the "Yes" moment which is a positive emotion that students express after solving a problem. The data were gathered while the participants were interacting with an EA-coach (an intelligent tutoring system that instructs introductory-level physics in the form of problem-solving scenarios). The researchers defined a novel set of features related to the "Yes moment", which included temporal information, sensory features, and pupillary response. The temporal information included the time students spent in answering a question, the number of attempts to produce the correct answer, and the degree of reasoning. The sensory features included: body position, skin conductance, mouse pressure, and pupil size.

Fifteen participants were asked to solve two physics problems involving at least 15 steps. During the session, the system recorded dialogue, temporal, and sensory information. Additionally, the participants were asked to verbally report their current emotions. The result of evaluating the full model (temporal information and the sensory information) and the temporal model demonstrated that the accuracy of the dialogue-based model was 81.6% and the full model was 81.4%. This result suggested that temporal information was sufficient to detect the "Yes" moment.

The researchers concluded that a dialogue-based features classifier could effectively distinguish moments of delight from other emotions better than models that use more complicated sensory features. Unlike D'Mello et al.'s study [3][9], the results of this study proved that moments of delight are successfully distinguishable using dialogue features. On the basis of the success of using dialogue-based features to detect moments of delight, we believe there is a great potential to successfully diagnose multiple emotions using dialogue-based features solely.

2.2.3.3 Overlapped-Emotions Recognition

Some emotions theorists believe that individuals may experience multi-overlapping emotions. Conati and Maclaren's [5] studied the effectiveness of using causal information to predict users' emotions. Two models were built, the first using causal information alone and the other using both diagnostic features and causal information. The diagnostic features were collected using an electromyography sensor (EMG) which monitors individuals' forehead muscles to diagnose their affect valence (negative/positive emotions). The model also used a goal-assessment subnetwork to infer the users' goals from their interaction patterns and traits. The affective user model in this study aimed to diagnose five emotions: emotions users experience during their interaction with video games (pride/shame), emo-

tions the user feels toward the feedback messages (admiration/reproach), and emotions the user feels toward the game (joy/regret).

Both models were evaluated on two datasets: a clear-valence data set where students clearly stated their emotions and an ambiguous-valence data set where students reported conflicting emotions. The combined models were significantly better than the causal model on the first data set, but for the second dataset the combined models were significantly less accurate than the causal model. The difference could be attributed to the measurement of conflicting emotions being less detectable using the forehead sensor, thus leading to less accurate sensory information.

Despite the somewhat discouraging results, this study is significant as the first work to detect multiple overlapping emotions during interaction with educational games. The model in this study also assessed users' emotions toward both game and system feedback. It is also the only study so far to take into account the causes of emotional change.

Chapter 3

Methodology

This chapter consists of two main sections. The first section will describe the architecture and the components of the intelligent tutoring system used in this study to collect interaction data. The second section will review the field study that was conducted to gather users' interaction data.

To build emotionally intelligent classifiers, researchers in affective computing typically conduct field studies to collect users' behavioural or physiological patterns which are then mapped to emotion categories. They either induce participants' emotions in a laboratory setting using one of the Mood Induction Procedures (MIPs) e.g., film or story, or in a real-world setting in which participants use their personal computers in their daily lives [40]. Both approaches have advantages and disadvantages: a laboratory-based study using mood induction procedures will yield more cleanly labelled data. However, the induced emotions do not necessarily represent the emotions that users experience in the real world.

On the other hand, using a real-world approach generates a greater amount of data compared to a time-limited laboratory-based approach, but with more noise and more incomplete data points. In this study, we chose to gather spontaneously generated interaction data without using any Mood Induction Procedures (MIPs), using same laboratory setting, computer application, and computer settings. The next section describes the components of the tutoring system used in this study.

3.1 CompTutor's architecture and design

For the purpose of collecting participants' interaction data, we built CompTutor, a dialogue-based tutoring system that teaches computer-related topics through a written dialogue. As shown in Figure 3.1, CompTutor consists of a student model, domain model, feedback generator, diagnoses module, and features extractor.

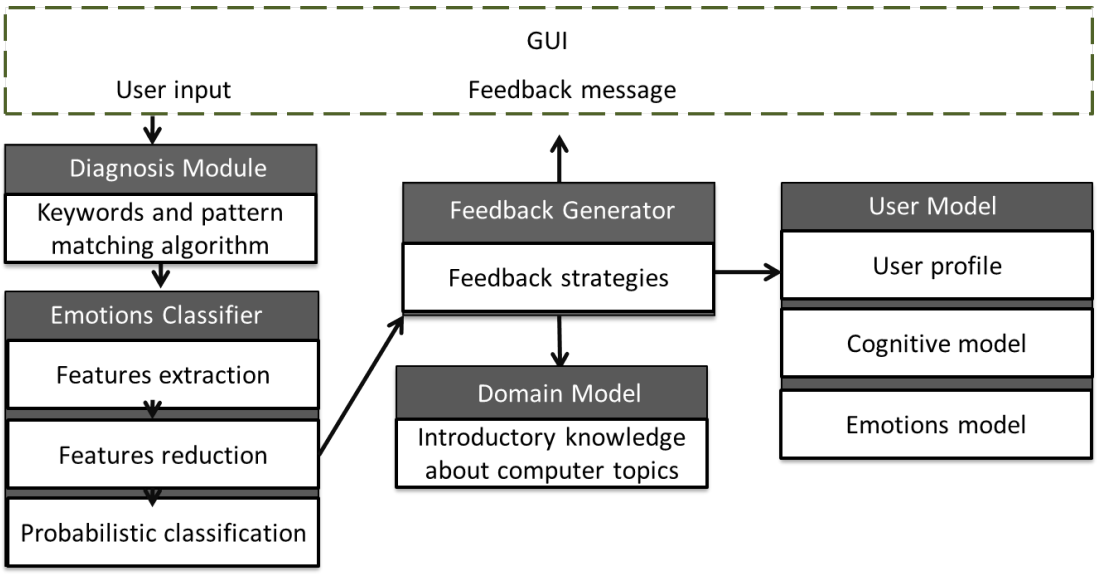

Figure 3.1: The architecture of CompTutor

3.1.1 The User Model

The user model is the component of the system that keeps track of the student's current state of knowledge. This is used to better understand the users' level of understanding and to adjust the tutoring accordingly. The user model in this study consists of a student profile, student cognitive model, and student affect model.

3.1.1.1 The User Profile

The user profile as shown in Figure 3.2 consists of the user's demographic information including name, age, gender, first language, educational level, motivation, and a general

description of the student's level of knowledge about different computer topics. This information is entered by the user at the start of using the system. The information in the user profile is stable and does not need to be updated other than the user's level. The user's reported level is used as an initial assumption of the user's level of understanding, and is calculated after each interaction made by the user.

The users' demographic information are: educational level, first language, age, and motivation. We took this information into consideration when evaluating users' typing behaviour as it might have an influence in users' typing behaviour. Individual tend to type faster and more accurately when they type in their native language. Users' age and educational level also have been proven to influence their typing behaviour [15] [33]. Also, users' motivations or reasons for using the systems could influence their emotions and typing behaviours. Users who use the system for study or work are more enthusiastic about using the system and try to answer the questions correctly; however, those who are only exploring the system are less enthusiastic. In this study, we did not include user's motivation in our analysis as all the subjects were using the system to participate in the study.

3.1.1.2 The Cognitive Model

After each interaction, the system evaluates the user's level of understanding in each topic by considering the user's belief about their level and by calculating the user's correct and incorrect answers. The student's level is categorized into beginner, intermediate, or advanced.

3.1.1.3 The Emotional Model

The emotional model includes the students' current emotions and the previous emotions inferred from the users' interaction. As shown in Table 3.1, the student information in this model is saved in the database and updated after every interaction made by the user.

3.1.2 Domain Model

The domain model, or the expert model stores all domain knowledge of CompTutor. As shown in Figure 3.3. The domain model consists of four topics: Foundation of Information Technology, Computer Hardware, Computer Software, and Internet Technologies. Each topic has 21 questions divided into three levels: beginner, intermediate, or advanced. The

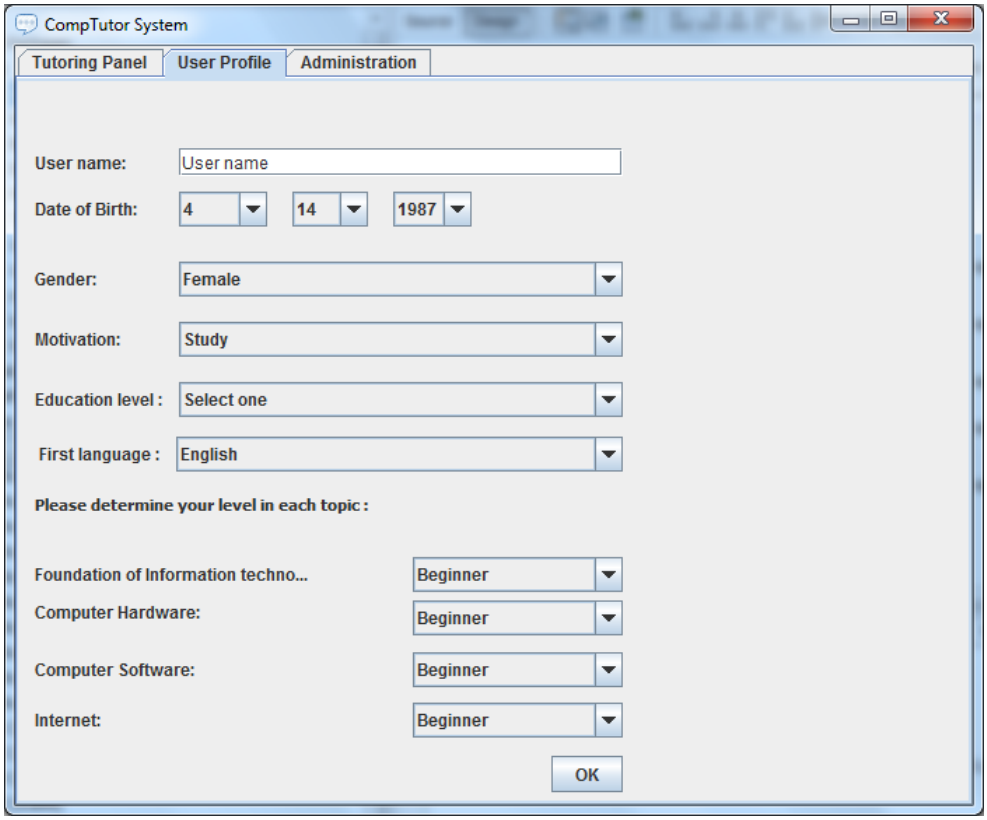

Figure 3.2: The user profile

questions as shown in Appendix B, were collected from different international computing tests such as IC3 (Internet and Computing Core Certification), ICDL (International Computer Driving License) and ECDL (European Computer Driving Licence). However, the questions were changed into an open-ended form to encourage students to write longer sentences. After choosing the topic from the main screen, the system randomly asks the user questions according to his/her level, taking into account the previously asked questions.

3.1.3 Feedback Generator

The feedback generator is the component that stores the tutoring and feedback strategies. After evaluating the user's responses, the system provides immediate feedback which belongs to one of the categories below.

Type	Items	Form	Acquisition Method	Acquisition Time
User Profile	User name	Text	User Input	First Time
	Date of Birth	Integer	User Input	First Time
	Gender	Integer	User Input	First Time
	First language	Integer	User Input	First Time
	Education level	Integer	User Input	First Time
	Motivation	Integer	User Input	First Time
Cognitive Model	User level in Topic 1	Integer	User input/ Calculated	Continuous
	User level in Topic 2	Integer	User input/Calculated	Continuous
	User level in Topic 3	Integer	User input/Calculated	Continuous
	User level in Topic 4	Integer	User input/Calculated	Continuous
Emotional Model	Emotions pattern	Array	Inferred	Continuous

Table 3.1: The components of the user model

- If the answer is correct and the user has not reached the maximum number of tries, the system randomly chooses a positive feedback (e.g., "Exactly", "That's Right") accompanied by a restatement of the correct answer.
- If the answer is incorrect and the user has not reached the maximum number of tries, the system asks the user to try again (e.g., "Incorrect answer, try again").
- If a part of the answer is correct and the user has not reached the maximum number of tries, the system asks the student to elaborate (e.g.,"Can you elaborate?", "What else?").
- If the answer is incorrect and the user has reached the maximum number of tries, the system states that the response is incorrect and gives the correct answer.

3.1.4 Diagnosis Module

The diagnosis module simply determines whether the user's response is correct, partially correct, or incorrect by checking the response for keywords related to the possible correct answers. If the module does not find any of the keywords, the response is considered a wrong answer. However, if the module finds part of the correct answer, the system considers it a partially correct answer.

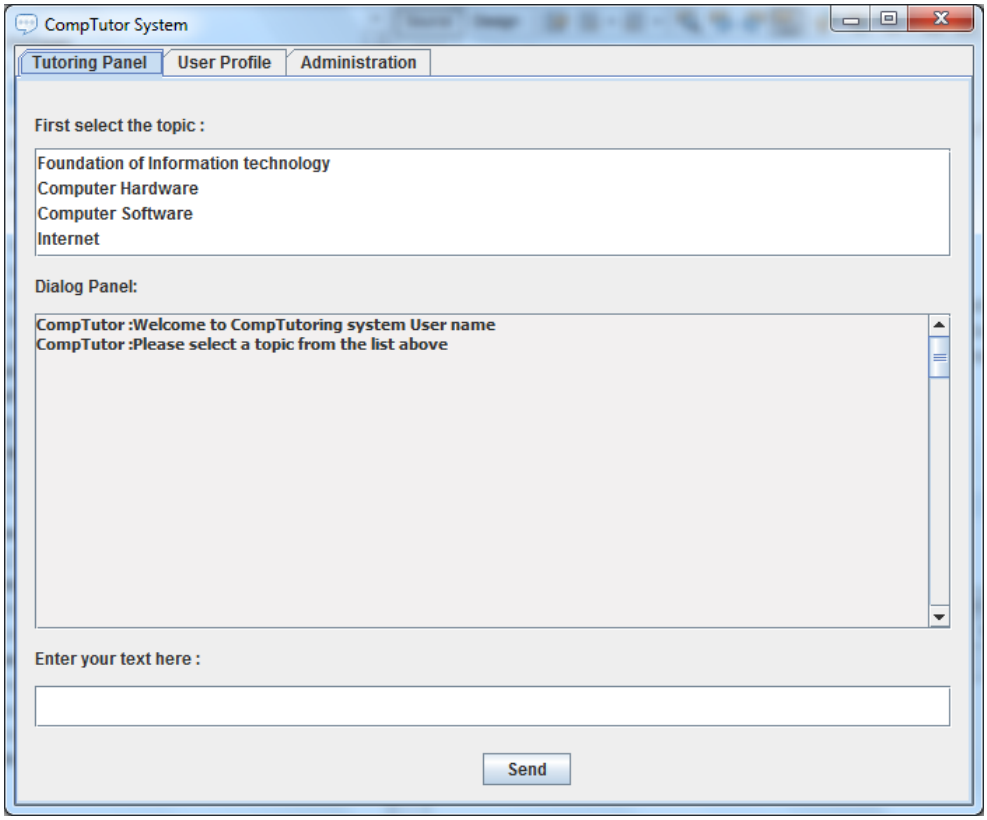

Figure 3.3: The main screen of CompTutor

3.1.5 Emotions Classifier

After each interaction, each user was asked to determine his/her emotions through a short questionnaire which asks participants to agree or disagree with five statements using a five-point Likert scale as shown in Figure 3.4.

After mapping the extracted features to the users' emotions, a hierarchical classifier (see Figure 3.5) was built to first determine the emotional valence using a binary classifier, followed by a finer classification to determine what kind of positive or negative emotions were experienced. The classifier assessed five emotions: confusion, delight, boredom, frustration, and neutral. These emotions were chosen for this study because of their proven relation to the learning process [9][23][17][29] and because they have proven to be the most frequent emotions experienced during learning sessions.

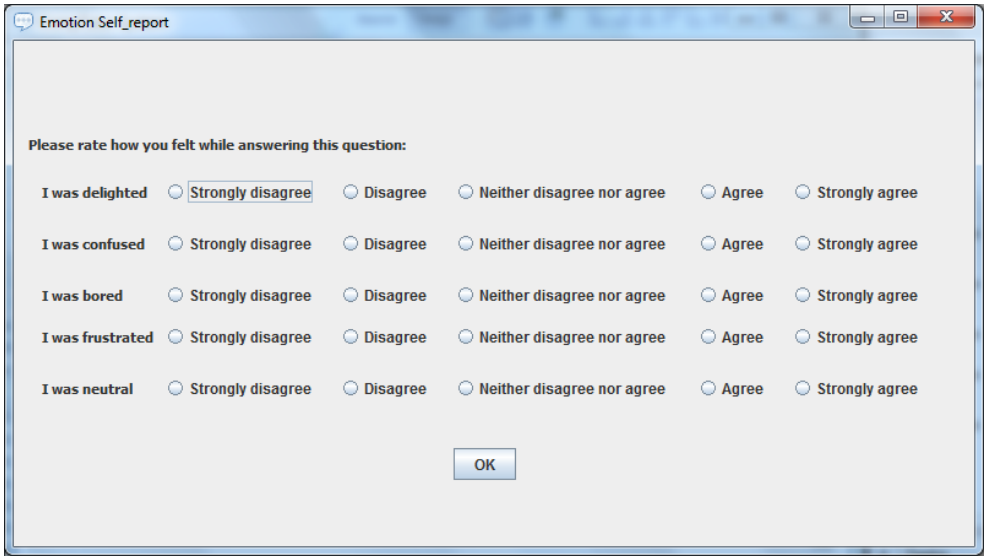

Figure 3.4: The user self-report short survey

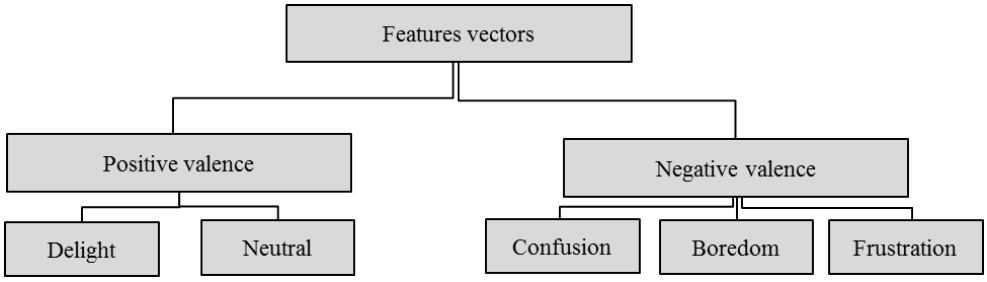

Figure 3.5: The architecture of the hierarchical classifier

3.1.6 Keystroke Logger

The CompTutor key logger calculates and extracts users' interaction data into an Excel file, which is then used to train and test the emotion classifier. After each user interaction, the system automatically calculates timing and keystroke features. As shown in Table 3.2, the focus of this study is on eighteen keystroke features commonly used in keystroke dynamics, stress detection, and affect detection studies. Keystroke dynamic studies generally capture two attributes: timestamps and key-codes of each depressed key by the user. The next section provides a brief description about each keystroke feature used in the study.

The mind data	Descriptions	Type	Acquisition method
Timing Features	Session duration	Continuous	Calculated
	Pause rate	Continuous	Calculated
Typing Features	Typing speed rate	Continuous	Calculated
	Key latency	Continuous	Calculated
	Key duration	Continuous	Calculated
	Deletion rate	Continuous	Calculated
	Capitalization rate	Continuous	Calculated
	Spaces per response	Continuous	Calculated
	Punctuation rate	Continuous	Calculated
	Unrelated keys rate	Continuous	Calculated
Response Features	Response quality	Discrete	Evaluated
	Response correctness	Discrete	Evaluated
	Number of words	Continuous	Calculated
	Spelling mistakes	Discrete	Evaluated
	Attempts per question	Continuous	Calculated

Table 3.2: The classification features

3.1.6.1 Timing and Keystrokes features

The features in this study consist of two values: key code and key timestamps which then used for further computation to generate timing and typing features described below :

- Session Duration: The duration that the user spends using the system. Session duration is calculated by computing the difference between starting time and user response times. Knowing how long the user has been cognitively stressed could help predict the state of boredom or frustration. The longer the user spent working on the system, the more likely they were to be in a negative emotional state.

- Typing Speed: As shown in Figure 3.6, the user typing speed is computed by taking the average of the time between the depression of a key and depression of the following key for each key pressed per response. The typing speed could be an indicator of different emotions [36].

- Deletion Rate: Deletion rate is the rate of using the backspace key and delete key per response. This rate may provide evidence of a state of confusion, as individuals who tend to delete are more likely to be confused [36].

- Key Latency: As shown in Figure 3.6, key latency is the time difference between releasing a key and depressing another key.
- Key Duration: As shown in Figure 3.6, key duration refers to the key depression time, which is computed by averaging the time from depressing a key to releasing it.
- Pause Rate: The time the user spends responding to a question, in other words the time difference between the question time and the response time. Spending more time answering a question could indicate confusion or boredom, whereas spending less time could indicate delight or neutral feelings [36].
- Use of Unrelated Key Rate: The rate of pressing non-letter keys such as numbers, arrows, or function keys. Using unrelated keys could be an indication of cognitive stress, which helps to predict the state of frustration or confusion.
- Capitalization Rate: The rate of using the Shift + letter or caps locks key.
- Punctuation Rate: The rate of using punctuation keys, including the semicolon, comma, dash, and period. Using capitalization and punctuations marks indicates that the user is typing carefully, which could be an indication of experiencing positive feelings.

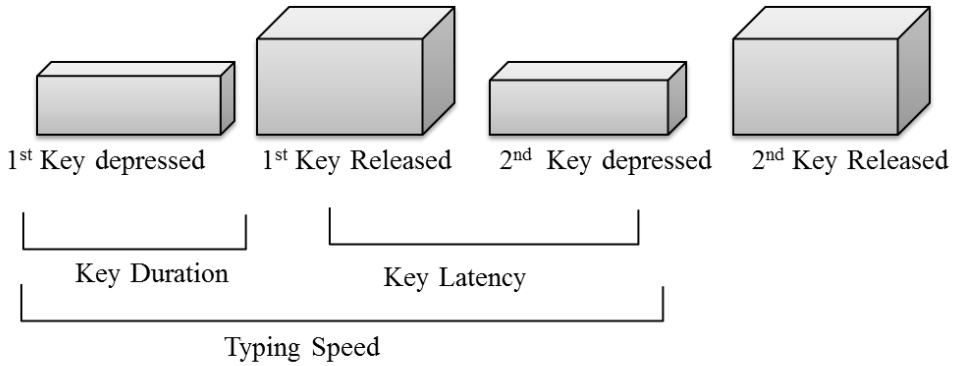

Figure 3.6: The key duration, key latency, and typing speed

3.1.6.2 The Response Quality

As shown in Table 3.2, the quality of users' responses were evaluated by a human judge using ordinal numbers. The response features include correctness, spelling mistakes, and

the quality of the users' answers. The correctness has four possible ratings: unrelated, incorrect, incomplete, and complete response. An unrelated response is the response that is not related to the general purpose of the system such as gaming the system. The quality of a response is rated in one of these categories: words, run-on sentence, full sentence, and well written sentence. The term spelling mistake represents the number of spelling errors in the users' response.

3.2 Field Study

Our study is concerned with diagnosing users' affective states during their interaction with tutoring systems. This study was conducted using CompTutor, a Windows application we developed using Java. The system teaches computer-related topics and gathers users' keystroke features. The interaction data includes the users' keystroke and timing features, response quality, and subjective and objective rating of the participants' emotions. The data was gathered during the participants' ongoing interaction, and no additional task was required of the user.

3.2.1 Getting Started

The participants were recruited through emails and posters (see Appendix A). Several emails were sent to undergraduate and graduate students' mailing lists in different departments of the University of Waterloo. Twenty participants took part in the field study, which was held at the University of Waterloo from January 27, 2011 to February 15, 2011. Before starting the experiment, the participants were required to sign a consent form describing the study's goals and procedure, and were asked for permission to use their interaction data and their video recordings.

To avoid discrepancies and possible questionable results, each participant in the study used the same application, tools, and laboratory setting. This avoided use of different types of keyboard or having different keyboard settings which could have otherwise affected the results [37]. For all subjects, the study was conducted in the same room. Subjects also used the same laptop, operating system (Windows 7), and application (CompTutor). In addition, keyboard settings were the same for all subjects, with a repeat speed and repeat delay of 15 and 2, respectively.

3.2.2 Study Procedure

After the procedure was explained and participants' inquiries were answered, participants were asked to use an installed version of the system on the researcher's laptop. The participants used the system without any supervision by the researcher, to avoid influencing their emotions. The field study consisted of two sessions. In the first session, participants were asked to type part of a paragraph that consisted of around 100 words. Their typing features were then extracted. This session was aimed at evaluating users' typing abilities in a neutral context, thereby providing a baseline for the classification model.

In the second session, each participant was asked to use CompTutor for around 45 minutes. First, they were asked to fill out their user profile with their information, including name, age, gender, educational level, and cognitive level on different computer topics as shown in Figure 3.3. Then, they were asked to begin the tutoring session. Both their behavioural responses and their screens were recorded using Camtasia Studio, a video and screen recorder.

Participants began the tutoring session by choosing one of the system topics from the main screen: information technology, computer hardware, computer software, and internet technologies. After choosing a topic, the system asked the participants a question according to their level and gave them three tries to provide a correct answer. After providing the correct answer or reaching the maximum number of tries, the system restated the correct answer and asked the user to select another topic.

After each response, the system computed the participants' typing features and asked each user to determine his/her emotion using a short questionnaire. The short questionnaire asked the participants to agree or disagree on five statements using a five-point Likert scale as shown in Figure 3.4.

3.2.3 Study Completion

After completing the experiment, participants were remunerated for their participation ($10/ hour), and were given a feedback letter. The quality of users' responses was evaluated by a human judge who evaluated the correctness and the quality of the answers according to the description of the ratings mentioned above.

Two psychology PhD students who had been trained in emotional intelligence evaluated each participant's emotions according to his/her recorded video. The observers were instructed to rate only the prominent emotion in case of ambiguous or overlapped emo-

tions. Both observers' affect rating and self-rating were used to train and test the affect classifier.

3.2.4 Participants' Demographics

Twenty participants from the University of Waterloo took part in this study: eight undergraduate and twelve graduate students, made up of nine females and eleven males. The participants were from different faculties: engineering, mathematics, art, and science. Their ages ranged from 19 to 34, with an average age of 24. Six of the participants indicated that their first language was English and 14 participants indicated that their first language was not English but rather Chinese, Arabic, Urdu, or Hindi.

Chapter 4

Data Collection

This chapter consists of two main sections. The first section will describe the nature of the dataset and review the data preparation procedures including data normalization and sampling. The second section will review the classification methods used in this study and the results of each method.

4.1 Data Preparation

Three datasets were used in this study: the user-labelled, judge1, and judge2 datasets. On average, each participant provided around 30 data points in half an hour, which were then associated with one of five emotion categories (delight, neutral, confusion, boredom, and frustration), and one of the emotional valence groups (positive valence and negative valence). The subjects' responses ranged from one word to 103 words, where the responses average was five words.

Originally the dataset consisted of 662 features vectors. However, the user-labelled dataset was reduced into 544 data points after excluding the incomplete and heterogeneously rated data points. An incomplete vector refers to a data point that does not have values in some of the features. Heterogeneous ratings refers to the data points that have contradictory labels such as choosing delight and bored. The data of the two judges was reduced into 581 data points after excluding incomplete data points. The next section will discuss the data filtering procedure in detail.

4.1.1 Data Filtering

During the self-labelling procedure, participants were instructed to rate their emotions using a five-point scale, but only a few participants used the scale properly. Few participants answered all five questions in the survey. Therefore, only the values of strongly agree and agree were combined and used, and the rest of the options were excluded.

During the experiment, subjects were asked to choose the dominant emotion that they felt. In this dataset, 41% feature vectors were associated with multiple ratings. Eight percent were associated with a contradictory rating, and 91% were within the same valence group (positive/negative). In the case of having different labels that belonged to the same group, the affective state with the highest rate in the scale was chosen. The features vectors that had contradictory labels belonging to different emotional valences were excluded from the study.

After filtering the data, the user-labelled dataset consisted of 164, 80, 146, 77, and 77 data points in each class of delighted, neutral, confused, bored, and frustrated, respectively. In total, 244 data points were labelled as positive emotions and 300 data points were labelled as negative emotions.

As shown in Table 4.1, the first judge dataset consisted of 90, 162, 142, 124, and 63 data points in each class of delighted, neutral, confused, bored, and frustrated, respectively. In total, 252 data points were labelled as positive emotions and 329 data points were labelled as negative emotions.

The second judge dataset consisted of 27, 164, 135, 168, and 87 data points in each class of delighted, neutral, confused, bored, and frustrated, respectively. In total, 191 data points were labelled as positive emotions and 390 data points were labelled as negative emotions.

4.1.2 Features Extraction

As explained in the previous chapter, the key logger computes and extracts all the typing features. Table 4.2 presents a list of all the typing features that were used in this study. According to the central limit theorem, we found that most of the features were normally distributed in each class except for those with ordinal discrete values: first language, educational level, answer quality, and answer correctness.

The dependency between the features was computed by Pearson correlation coefficient and indicated that most of the features were independent or had low correlation with each

Emotional Valence	Emotional States	User-labelled	Judge1	Judge2
Positive	Delighted	164	90	27
	Neutral	80	162	164
	Total Positive	244	252	191
Negative	Confused	146	142	135
	Bored	77	124	168
	Frustrated	77	63	87
	Total Negative	300	329	390
Total data points		544	581	581

Table 4.1: The number of data points in each dataset

other, while a few were highly correlated or moderately correlated with others. The user typing speed and key latency were significantly correlated with $r = 0.940$ and $p<0.01$. The correlation between typing speed and key latency was attributed to the similarity between those features, where typing speed was the difference between depressing a key and depressing another key, and key latency is the difference between releasing a key and depressing another key.

The deletion rate was also significantly correlated with the rate of using punctuation marks, the rate of pausing, and the length of response with $r = 0.587$ and $p<0.01$, $r = 0.581$ and $p<0.01$, and $r = 0.545$ and $p<0.01$, respectively. The length of answers is moderately correlated with answer quality, spelling mistakes, and correctness with $r = 0.516$ and $p<0.01$, $r = 0.505$ and $p<0.01$, and $r = 0.499$ and p<0.01, respectively.

Classification Features	**Description**
Session Duration	Session starting time − User response time
First Language	English 1, other 2
Educational Level	Undergraduate 1, Graduate 2
Pause Rate	System response time − User response time
Typing Speed	$\frac{\sum \text{Timestamp of Key1}-\text{Timestamp of Key2}}{\text{Characters per sentence}}$
Deletion Rate	$\frac{\sum \text{Use of backspace key}}{\text{Characters per sentence}}$
Use of Unrelated Keys	$\frac{\sum \text{Use of non-letters keys}}{\text{Characters per sentence}}$
Numbers of Tries	Ordinal number range from 0 to 3
Length of Response	Number of characters per response
Use of Punctuation Marks	$\frac{\sum \text{Use of punctuation marks}}{\text{Characters per answer}}$
Use of Spaces	Number of spaces per response
Capitalization Rate	$\frac{\sum \text{Use of Shift and Letter}+\text{Use of Caps locks}}{\text{Characters per sentence}}$
Key Duration	$\frac{\sum \text{Depressing timestamp of Key1}-\text{Releasing timestamp of Key1}}{\text{Character per sentence}}$
Key Latency	$\frac{\sum \text{Releasing timestamp of Key1}-\text{Depressing timestamp of Key2}}{\text{Character per sentence}}$
Answer Quality	Range from 1: words to 4 : grammatically correct answer
Spelling Mistakes	$\frac{\sum \text{Spelling mistakes}}{\text{Character per sentence}}$
Answer Correctness	Range from 1: unrelated answer to 4 : correct and complete answer
User level	Ordinal number 1:beginner, 2:intermediate, and 3: advanced

Table 4.2: Description of classification features

4.1.3 Data Normalization

One of the big challenges in the user modeling and pattern recognition areas is whether the relationship between the user's pattern and the labels can be generalized across different individuals or not. Therefore, the interaction features in this study were normalized according to each participants' typing behaviour. As explained in Chapter 3, the participants provided the interaction data under two conditions: a baseline condition, and an experimental condition. In the first condition, subjects typed a part of a paragraph in a neutral context which was used as a baseline to normalize the data. The data was normalized by subtracting from each feature value by the mean value of the feature and dividing by the standard deviation value.

The mean and the variance were computed using the users' typing features under the baseline condition. All of the classification methods were performed on both datasets: the raw dataset, and the normalized dataset. Normalizing the typing features according to each user's typing behaviour improved affect detection in past studies [38]. In this study, the classification of the raw data was slightly more accurate than the normalized typing features, as will be discussed later, except using Linear Discriminant Analysis (LDA) and k-nearest neighbour.

4.1.4 Sampling Methods

To estimate the error rate of the classifiers, two approaches were used: split-sample and cross-validation. Both sampling methods were utilized to train and evaluate all of the six classifiers except the neural network classifier where only the split-sampling was used. Single factor ANOVA was performed to compute the difference between the classification rate of using split-sampling and cross-validation, and indicated no significant difference between these methods with $MS = 0.0002$, $F = 0.009$, and $p>0.05$.

In the split-sampling method, the data was randomly divided into two sets each consisting of 70% and 30% of the dataset for training and evaluation, respectively. The training and the testing sets each had 70% and 30% of the observations from each class. The cross-validation method performed using MATLAB standard five-fold cross validation function which randomly divides the data into five disjoint stratified subsets. Four subsets were used for training and one for evaluation. This method generates two sets consisting of 80% and 20% of the original data for training and testing, respectively.

4.1.5 Features Reduction (Principal Component Analysis)

Principal Component Analysis is a statistical algorithm used to reduce the dimensionality of features by computing the eigenvalues and eigenvectors of the data and then finding features that account for the variation in data [41]. Although the dataset consisted of only 18 features, which is low compared with image processing and bioinformatics datasets, using PCA can provide more accurate classification by excluding the unrelated features. PCA was preformed on both the raw and the normalized dataset of the three datasets: users-labelled, judge1, and judge2 datasets.

The result of applying PCA on the normalized judges' dataset demonstrated that eight to ten eigenvector were responsible for 77% of the variance, while only three eigenvectors accounted for 95% of the variance of the non-normalized judges' dataset. However, using the extracted features of the normalized datasets yielded a significantly improved classification rate compare to the non-normalized dataset.

Applying the dimensionality reduction analysis on the normalized user-labelled dataset indicated that eight to ten eigenvectors accounted for 80% of the variance, while applying it on the non-normalized dataset indicated that three eigenvectors accounted for 96% of the variance in the data. However, in comparison with the non-normalized dataset, the extracted features of the normalized datasets yielded a better classification rate.

Six classification techniques were used to classify the extracted features using PCA and the raw datasets. Using the raw data provided more accurate classification results compared to using the decorrelated features. This results could be attributed to having only 18 features each is moderately accounting for some of the variation in the data.

4.2 Data Analysis

4.2.1 Relating Affect and Typing Features

Correlation analysis was implemented to understand the relationship between the interaction features and emotional change. The results of the correlation analysis confirmed the hypothesis of this thesis that typing features moderately yield viable information for affect detection. Pearson correlation was performed on all the features in the three datasets, where emotion categories were in the form of ordinal numbers ranging from delighted (1) to frustrated (5) and emotional valence ratings were positive (1) and negative (0).

The results of the correlation analysis in general did not show any significant correlation between features and user's emotion, but it only showed moderate to weak correlation with some features. The number of tries to answer a question had moderate-positive correlation with users' emotions in all the three datasets with $r = 0.310$ and $p<0.01$, $r = 0.399$ and $p<0.01$, and $r = 0.308$ and $p<0.01$ for the user-labelled, judge1, and judge2 datasets, respectively. These results showed that as the number of the attempts increased, the participants were more likely to be experiencing negative feelings.

Session duration and pause rate were also moderately correlated with emotions in the second judge dataset with $r = 0.245$, $p<0.01$ and $r = -0.213$, $p<0.01$, respectively. Session duration was positively correlated with emotions' values. The longer the session, the more likely the user was experiencing negative emotions. The pause rate was correlated negatively with users' emotions. The longer the pause rate, the more likely the user was experiencing positive emotions.

The length of response had a moderately negative correlation with emotions with $r = -0.230$ and $p<0.01$, $r = -0.161$ and $p<0.01$, and $r = -0.167$ and $p<0.01$ for the user-labelled, judge1, and judge2 datasets, respectively. Similarly, the answer correctness was also moderately correlated with emotions in the user-labelled, judge1, and judge2 datasets with $r = -0.424$ and $p<0.01$, $r = -0.239$ and $p<0.01$, and $r = -0.265$ and $p<0.01$, respectively. These results indicated that the more correct and longer the answer, the more likely the user is experiencing positive emotions.

The correlation analysis showed that deletion rate has some correlation with emotions in all the three datasets with $r = -0.103$ and $p<0.01$, $r = -0.172$ and $p<0.01$, and $r = -0.190$ and $p<0.01$. The more the user deleted, the more likely the user was in a positive affective state. On the other hand, the correlation analysis showed weak correlations between typing speed, key latency, key duration, and emotions in all the datasets.

Additionally, another Pearson correlation was computed on the data where the ratings of users' emotions were binary values of presence (1) or absence (0) of each emotion (see Table 4.3). This approach has been used previously by D'Mello et al [8] to analyze the relationship between emotions and dialogue features. The results of our analysis showed that some features were moderately correlated with some emotions categories. Table 4.3 includes any moderate correlation where $r>0.2$ and $p<0.01$. In the table the positivity and negativity of a correlation is presented by + and −, respectively.

On the basis of this analysis, we found that being delighted was associated with short response rate, use of deletion, fewer attempts, use of punctuation marks, use of capitalization, longer response, and correct answers. Confusion was associated with short session duration, long response rate, fast typing, long key latency, more attempts, short responses,

and incorrect answers. Boredom was associated with long session duration, short pause rate, and use of unrelated keys.

While the results showed no significant correlation between the state of being neutral and typing features, being frustrated was associated with long pause rate, less deletion rate, more attempts, less use of punctuation, less use of capitalization, short responses, incorrect answer and poorly written answer.

	User-labelled					**Judge1**					**Judge2**				
Feature	D	N	C	B	F	D	N	C	B	F	D	N	C	B	F
SD	+		-	+				-	+		-	-	-	+	
PR	-		+			-	-	+		+		-	+	-	+
DR	+	+	-	-	-					-			+		-
TS								+					+	-	-
KD	+	-													
KL			+					+					+	-	
AT	-		+	+	+	-	-	+		+		-	+	-	+
URK		+		+					+	-				+	-
UPM	+		-		-	+				-		+	-		-
UCL	+		-	-	-									+	-
LOR	+		-	-	-		+	-		-		+	-	+	-
AQ	+		-	-	-	+		-		-		+	-	+	-
AC	+		-	-	-	+	+	-		-		+	-	+	-

Table 4.3: Correlation between classification features and users' emotions
Emotions: D: delight, N: neutral, C: confusion, B: bored, F: frustrated. Features: SD: session duration, PR: Pause rate, DR: deletion rate, TS: typing speed, KD: key duration, KL: key latency, AT: number of attempts, URK: unrelated keys, UPM: use of punctuation marks, UCL: use of the capital letters, LOR: length of response, AQ: Answer quality, AC: answer correctness

4.2.2 Judgement Reliability Evaluation

To evaluate the emotion judgement agreement, Cohen's Kappa inter-rater agreement was computed for each pair of the judges: user-judge1, user-judge2, and judge1-judge2 for rating both emotions and emotional valence. The inter-rater reliability for the raters on evaluating users' emotion was found to be moderate with k=0.29 for user-judge1, k=0.19

for user-judge2, and k=0.32 for judge1-judge2 agreement. The results of the inter-rater agreement analysis indicated that the highest agreement was between the external judges in rating emotions. This results is consistent with D'Mello et al's [8] findings that trained judge's agreement score was highest compared to the the agreement between the other pairs.

While the inter-rater agreement scores for rating emotional states were moderate, the inter-rater agreement scores for rating the emotional valence were significantly higher than rating emotions. The Kappa scores for rating the emotional valence were k=0.52 for user-judge1, k=0.24 for user-judge2, and k=0.43 for judge1-judge2 agreement. On the basis of these results we found that judges may have a different interpretation in evaluating users' emotions, but they are more able to successfully agree on which emotion the user experienced in a more general sense (positive or negative).

Among all three pairs, user-judge1, user-judge2, and judge1-judge2 the Kappa scores in rating confusion (k=0.53) was the highest compared to boredom (k=0.45), frustration (k=0.42), neutral (k=0.42), and delight (k=0.30). These results demonstrated that some emotions could be more easily distinguished through facial expression, such as confusion and boredom, compared to other emotions where neutral and delight were less easy.

4.3 Classification Methods Assessment

Several classification methods were evaluated in this study based either on their previous use in affect detection studies or on outperforming other classification methods across a variety of other machine learning application. Discriminant Analysis, Bayesian Analysis, k-Nearest Neighbour, and Decision Trees are the most commonly used methods in affect recognition and keystroke dynamics studies. To the best of our knowledge, Artificial Neural Networks (ANN) have not previously been employed to detect emotion through typing behaviour.

Using MATLAB, these classification methods were assessed to find the classifier that outperformed the others. These methods were applied on three datasets (user-labelled, judge1, and judge2) in addition to the normalized dataset and the dimensionally reduced dataset. Tables 4.1, 4.2, and 4.3 report the classification accuracy of each classification method for raw data, normalized, and the dimensionally reduced dataset, respectively. The next section gives a brief description about each classification method used in this study.

4.3.1 Discriminant Analysis

Linear Discriminant Analysis (LDA) and PCA are both used for classification, but PCA is used to classify features by finding those most responsible for the variation in the data. LDA is used to separate the data according to their classes and draw a decision region. The main concept behind LDA is to separate the features by maximizing the space between the classes and minimizing the space within the class [24]. LDA was chosen for its simplicity and the capability of handling uneven amounts of observations in each class.

In this study, the three datasets were classified using the MATLAB discriminant function that computed using the pooled covariance matrix and the prior probability of each class. As shown in Table 4.4, the classification accuracy for classifying the emotional valence was 71.82%, 63.73%, and 69.43% for the user-labelled, judge1 and judge2 datasets, respectively. However, the accuracy of classifying users' emotions was 51.93%, 34.72%, and 44.04% for the user-labelled, judge1 and judge2 datasets which is considerably lower as the data is not linearly separable.

Similarly, Quadratic Discriminant Analysis(QDA) separates a set of observation data by a quadratic surface [24]. QDA works on the assumption that the data is normally distributed and there are only two classes to be distinguished from each other. The accuracy of classifying users' emotional valence in the user-labelled, judge1, and judge2 datasets as shown in Table 4.4 was 70.72%, 63.21%, and 61.66%, respectively. The accuracy of classifying users' emotions was also considerably low with 44.2%, 37.31%, and 46.55%, for user-labelled, judge1, and judge2 dataset, respectively.

4.3.2 Naive Bayes

Bayesian analysis is one of the simplest classification methods that computes the posterior probability of each class based on the observed patterns and choose the maximum posterior with the assumption of having independent features. Most of the features in the three datasets are normally distributed as discussed above. However, the Naive Bayes classifier was implemented using two assumptions: first that the data are normally distributed (Gaussian distribution), and second that the data are nearly normally distributed (Kernel distribution).

As shown in Table 4.4, using a Gaussian distribution assumption, the classification accuracy of emotional valence recognition was 71.27%, 60.10%, and 65.28% for the user-labelled, judge1, and judge2 datasets, respectively. Using the same assumption for classifying users'

emotions yielded an accuracy of 43.65%, 30.05% and 38.34% for the user-labelled, judge1 and judge2, respectively.

The classification accuracy for classifying emotional valence using the Kernel density function was 65.19%, 63.73%, and 70.98% for the user-labelled, judge1, and judge2 datasets, respectively. The classification accuracy of emotions recognition was 44.20%, 33.16%, and 40.41% for the user-labelled, judge1, and judge2 datasets, in turn.

The ANOVA test shows that there is no significant difference between using Gaussian or Kernel assumptions as both gave very close results with $F = 0.077$, $MS = 1.7604$, and $p>0.05$. However, the kernel-based classifier yielded a slightly better accuracy on both emotions and emotional valence classification.

4.3.3 k-Nearest Neighbour

K-nearest neighbour is a machine learning method that assigns features to the class of the nearest or the closest feature in the training set. Despite the simplicity of the KNN algorithm it is one of the most effective methods in pattern recognition. The classification accuracy using the MATLAB standard k-nearest neighbour algorithm with k=1 was 61.88%, 62.18%, and 63.21% for classifying the emotional valence of the user-labelled, judge1, and judge2 datasets, respectively. However, the accuracy of classifying emotional states was still low with 32.60%, 31.09%, and 37.82% for the user-labelled, judge1, and judge2 datasets, respectively.

4.3.4 Decision Trees

Decision Trees are a supervised machine learning technique used to classify a set of observations by generating a set of rules in a hierarchical structure. Several emotion-recognition studies have used decision trees to classify users' behaviour or physiological patterns. MATLAB standard regression tree was used to classify the feature vectors based on the emotional valence and emotional states in the three datasets.

The decision tree classifier in this study did not significantly outperform the other techniques. It yielded an accuracy of 67.96%, 65.28%, and 75.65% for the user-labelled, judge1, and judge2 datasets, respectively. And, the accuracy of emotions classification was 39.78%, 35.75%, and 41.45% for the user-labelled, judge1, and judge2 datasets.

4.3.5 Artificial Neural Networks

An Artificial Neural Network is a classification method that builds a network consisting of inputs, hidden layers, weighted edges, and outputs using a set of observations. The weight of the edges and the number of the hidden layers were chosen according to the optimal prediction results. Two standard feed-forward neural networks were built to determine emotional valence and emotional states.

The artificial neural network classifier yielded the best accuracy compared to the other classification methods. The classification accuracy yielded from the ANN emotional valence classifier was 82.82%, 72.02%, and 77.20% for the user-labelled, judge1, and judge2 datasets. ANN classifier was moderately able to differentiate between the six emotional categories with an accuracy of 53.59%, 45.60%, and 53.89% for the user-labelled, judge1, and judge 2 datasets.

Classification Methods	**Sampling Method**	**User-labelled**		**Judge1**		**Judge2**	
		Valence	Emotions	Valence	Emotions	Valence	Emotions
LDA	Split	71.82	51.93	63.73	34.72	69.43	44.04
	5-fold CV	72.61	50.37	64.31	33.97	70.00	49.66
QDA	Split	70.72	44.20	63.21	37.31	61.66	46.55
	5-fold CV	70.40	43.75	63.79	34.66	61.03	46.03
KNN	Split	61.88	32.60	62.18	31.09	63.21	37.82
	5-fold CV	61.76	38.42	57.07	31.38	60.69	36.90
DT	Split	67.96	39.78	65.28	35.75	75.65	41.45
	5-fold CV	68.20	47.24	62.93	44.14	69.83	41.03
NBG	Split	71.27	43.65	60.10	30.05	65.28	38.34
	5-fold CV	68.20	46.32	61.38	30.69	64.48	45.34
NBK	Split	65.19	44.20	63.73	33.16	70.98	40.41
	5-fold CV	66.36	43.57	61.72	30.69	67.41	45.17
ANN	Split	82.82	53.59	72.02	45.60	77.20	53.89

Table 4.4: The classification accuracy of the raw data
LDA: Linear discrimination analysis, QDA: Quadratic discrimination analysis, KNN: k-nearest neighbour, DT: Decision tree, NBG: Gaussian naive Bayes, NBK: Kernel naive Bayes, ANN: Artificial neural network

Classification Method	**Sampling Method**	**User-labelled**		**Judge1**		**Judge2**	
		Valence	Emotions	Valence	Emotions	Valence	Emotions
LDA	Split	73.48	52.49	69.95	40.41	72.02	45.60
	5-fold CV	73.53	50.55	64.37	34.94	70.74	45.78
QDA	Split	55.80	50.83	44.56	34.20	48.19	32.64
	5-fold CV	57.72	48.53	63.79	34.42	44.41	32.70
KNN	Split	71.27	49.72	61.66	35.23	67.36	44.04
	5-fold CV	72.61	47.24	60.71	35.28	65.75	45.09
DT	Split	68.51	43.65	65.28	38.67	66.85	45.86
	5-fold CV	67.28	45.59	62.93	36.66	69.19	43.89
NBG	Split	49.17	45.30	47.67	36.79	64.77	21.24
	5-fold CV	51.29	47.06	58.35	34.42	62.31	23.06
NBK	Split	63.54	48.62	58.55	35.75	66.84	40.41
	5-fold CV	62.32	48.62	61.79	35.75	66.61	41.44
ANN	Split	72.66	51.93	68.92	43.53	74.61	56.48

Table 4.5: The classification accuracy of the normalized dataset
LDA: Linear discrimination analysis, QDA: Quadratic discrimination analysis, KNN: k-nearest neighbour, DT: Decision tree, NBG: Gaussian naive Bayes, NBK: Kernel naive Bayes, ANN: Artificial neural network

Classification Method	**Sampling Method**	**User-labelled**		**Judge1**		**Judge2**	
		Valence	Emotions	Valence	Emotions	Valence	Emotions
LDA	Split	71.80	47.51	66.84	29.53	67.88	41.45
	5-fold CV	71.51	46.69	63.34	31.03	67.64	42.07
QDA	Split	67.40	46.96	66.84	26.94	65.28	20.73
	5-fold CV	70.40	47.79	64.72	27.24	70.40	39.48
KNN	Split	71.27	45.30	62.69	40.41	70.47	46.63
	5-fold CV	68.20	43.01	63.51	36.90	64.03	43.97
DT	Split	61.88	38.67	58.03	25.39	59.07	38.86
	5-fold CV	65.62	40.07	61.55	29.14	64.48	39.66
NBG	Split	71.27	43.65	62.69	34.20	64.77	33.16
	5-fold CV	67.28	44.67	62.07	31.03	63.10	34.10
NBK	Split	65.19	44.20	66.84	40.41	66.32	37.31
	5-fold CV	65.81	44.67	63.62	33.10	69.83	40.69
ANN	Split	69.54	42.19	66.32	37.30	70.98	43.52

Table 4.6: The classification accuracy of the dimensionally reduced dataset
LDA: Linear discrimination analysis, QDA: Quadratic discrimination analysis, KNN: k-nearest neighbour, DT: Decision tree, NBG: Gaussian naive Bayes, NBK: Kernel naive Bayes, ANN: Artificial neural network

4.3.6 Classification Accuracy Comparison

4.3.6.1 Comparison Across Normalized, Raw, and Decorrelated dataset

In general, normalizing the data according to each participants' typing behaviour yielded slightly better classification accuracy using classification methods such as k-nearest neighbour and linear discrimination analysis. An ANOVA was performed to compare the difference between the normalized and the raw datasets for both emotional valence and emotional states recognitions. The results of the analysis showed no significant difference between the classification of the normalized dataset and the raw dataset with $MS = 32.730$, $F = 3.125$, and $p>0.05$ for emotional valence recognition, and $MS = 3.763$, $F = 0.133$, and $p>0.05$ for emotional state recognitions. However, the raw data yielded a better classification for emotional valence while the normalized dataset yielded a slightly better accuracy for emotional state recognition.

By comparing the average accuracy of the dimensionally reduced dataset and the raw dataset across all the classification methods, we found no significant difference between the two datasets with $MS = 3.472$, $F = 0.502$, and $p>0.05$ and $MS = 10.594$, $F = 0.404$, and $p>0.05$, for classifying emotional valence and emotional states, respectively. The raw data provided more accurate classification for both emotional valence and emotional state recognition.

4.3.6.2 Comparison Across Affect Judges

One-way ANOVA was performed to measure the difference between the accuracy of classifying emotional valance among the three datasets. The results showed no significant difference in the classifications of emotional valence across the user-labelled, judges1, and judge2 datasets: $MS = 68.629$, $F = 2.191$, and $p>0.05$. However, ANOVA showed a significant difference of the classifications of emotions among the three datasets, $MS = 165.197$, $F = 4.558$, and $p<0.05$. Using Tukey's post-hoc test, we found that the user-labelled dataset provided the most accurate classification compared to the judges dataset on detecting both emotional valence and emotional states.

4.3.6.3 Comparison Across Classification Methods

One-way ANOVA revealed a statically significant difference in the classification accuracy between the six classifiers on detecting emotional valence with $MS = 68.419$, $F = 3.460$, and $p<0.05$, and also emotional states, $F = 3.295$, $MS = 90.832$, and $p<0.05$. Tukey's

post-hoc test indicated that ANN classifier generated the most accurate classification in comparison with the other methods in detecting both emotional valence and emotional states.

4.3.6.4 Comparison Across Affective States

By analyzing the confusion matrices of each classifier in diagnosing emotional valence and and emotional states, shown in Table 4.7 and 4.8 on all three datasets, we found that among all classifiers the positive emotions were less predictable compared to the negative emotions. Moreover, some classifiers were more able to classify certain emotions more accurately than the other methods. Noteworthy that the number of observations in each group was not equal as some datasets had a disparate number of observations in each class, e.g., judge2 dataset had only 27 feature vectors which were then divided into training and testing subsets.

Two-way ANOVA was performed on the classification accuracy of each emotion in the three datasets. In the user-labelled dataset, there was a significant difference between classification accuracy of different emotions with $MS = 1983.743$, $F = 16.968$, and $p<0.01$. in the user-labelled dataset, the most predictable emotions among all classifiers were delight and frustration, while neutral and bored were the least predictable emotions. In judge2 dataset, the ANOVA showed a significant difference of the classification accuracy across the five emotions with $MS = 2260$, $F = 20.782$, and $p<0.01$. The most predictable emotions in judge2 dataset were confusion, boredom, and frustration. Neutral was the least predictable emotion in judge2 dataset which was due to the low number of the data points associated with neutral. In contrast, there was no significant difference in classifying different emotions in the judge1 dataset $MS = 227$, $F = 0.829$, and $p>0.05$.

By analyzing the numbers of true negatives, false positives, false negatives, and true positives, we found that most of the misclassified vectors (false negative) of delight, boredom and confusion were often classified as neutral while most of the false negative values of the neutral vectors were classified as confusion or boredom. On the other hand, some classifiers misclassified some of the frustration vectors as confusion or boredom. On the basis of these results, we found that neutral is the most unrecognizable emotion compared with other emotions such as delight, boredom and confusion. This finding may be attributed to the nature of these emotions, which share similar characteristics.

Classification Methods	**User-labelled**		**Judge1**		**Judge2**	
	Positive	Negative	Positive	Negative	Positive	Negative
LDA	75	72	65	64	76	73
QDA	60	77	77	55	76	52
KNN	65	61	43	62	46	76
DT	77	68	61	67	65	78
NBG	53	76	72	63	63	62
NBK	48	76	38	76	46	79
ANN	72	72	71	77	72	77

Table 4.7: The classification rate per emotional valence (positive, negative)
LDA: Linear discrimination analysis, QDA: Quadratic discrimination analysis, KNN: k-nearest neighbour, DT: Decision tree, NBG: Gaussian naive Bayes, NBK: Kernel Naive Bayes, ANN: Artificial Neural network

Classification Methods	**User-labelled**					**Judge1**					**Judge2**				
	D	N	C	B	F	D	N	C	B	F	D	N	C	B	F
LDA	67	22	44	20	69	48	33	41	36	38	33	35	51	57	66
QDA	37	26	35	40	46	41	43	48	17	48	0	60	50	49	55
KNN	57	26	27	28	27	21	24	33	37	24	0	46	29	39	44
DT	60	22	47	20	69	17	39	42	41	22	11	40	49	50	37
NBG	43	18	29	20	62	20	37	17	15	71	0	56	30	36	66
NBK	57	11	35	16	69	6	40	31	18	71	25	42	48	42	66
ANN	70	33	59	20	60	85	41	46	48	46	0	58	48	60	57

Table 4.8: The classification accuracy per emotions
D: delight, N: neutral, C: confusion, B: boredom, F: frustration, LDA: Linear discrimination analysis, QDA: Quadratic discrimination analysis, KNN: k-nearest neighbour, DT: Decision tree, NBG: Gaussian Naive Bayes, NBK: Kernel naive Bayes, ANN: Artificial Neural networks

4.4 Hierarchical Classification

Among the six classification methods, the accuracy of determining users' emotional valence were significantly higher compared to determining emotional states. The inter-rater agreement scores were also significantly higher in rating users' emotional valence compared to their emotions. These results could be attributed to several factors such as the lower number of classes and the easiness of identifying emotional valence compared to identifying emotional states. Thus, to lower the classification error rate a hierarchical classification was implemented to first classify users' emotional valence and then determine the emotional state.

Hierarchical emotion classifiers have been proposed by several affect recognition researchers; e.g., Hoque et al. [13] and Lin [20]. The primary goal in classifying users' emotions into a hierarchy form is to reduce the number of classes in each step by first classifying the observed patterns into positive or negative emotional valence and then into a finer classification of what specific positive or negative emotion the user experienced.

In this study, the artificial neural network classifier was ultimately chosen as it yielded the best classification results in comparison with the other methods. Three feed-forward neural networks were built to first classify users' emotional valence in the three datasets. Then, two neural networks were trained on the positive and negative emotions separately for each of the three datasets. The neural network of positive emotions distinguished between two groups: delighted and neutral, while the second neural network distinguished between three categories: confusion, boredom, and frustration.

Classification Accuracy	**User-labelled**		**Judge1**		**Judge2**	
	Raw	Normalized	Raw	Normalized	Raw	Normalized
Emotional valence	82.82	72.66	72.02	68.92	77.20	74.61
Negative emotions	68.83	63.64	66.97	64.22	71.32	69.77
Positive emotions	80.24	82.36	73.81	75.00	89.06	92.19
Overall classification	59.37	63.28	49.74	44.04	56.48	55.44

Table 4.9: The classification accuracy of the hierarchical classifier for the user-labelled, judge1, judge2 datasets

Table 4.9 presents the classification accuracy of the emotional-valence classifier, negative-emotions classifier, and positive-emotions classifier. Similar to the standard classification methods, the classifications of the user-labelled dataset was better than the judges'

datasets. The results showed that the overall classification rate of the hierarchical classifier outperformed the standard approaches of classifying the five emotions together.

Using the hierarchy approach to classify users' emotions improved the classification accuracy. However, it moderately outperform the standard neural network classifier that distinguished between the five emotions with $MS = 26.083$, $F = 1.121$ and $p>0.05$ for the raw data and $MS = 19.512$, $F = 1.121$, and $p>0.05$ for the normalized data. Despite the moderate improvement, the overall accuracy obtained using the hierarchical classifier outperformed the previous dialogue-based affect recognition approaches and in line with some affect recognitions approaches that used expensive computations or additional tools, as will be discussed in Chapter 5 in detail.

Chapter 5

Discussion

This chapter reviews the goals and contributions of the research, summarizes the results presented in Chapter 4, discusses the limitations of this study, and introduces potential future work.

5.1 Research Overview

This research explored the correlation between typing features and users' emotions. The main goals of this research were building a dialogue-based tutoring system to collect interaction data, collecting interaction data associated with self and expert judgements, analyzing the correlation between typing features, emotional states, and emotional valence, evaluating different classification methods in diagnosing emotions, and implementing hierarchical classification to diagnose emotional valence and determine emotional state.

5.2 Evaluation of Judgement Agreement

The results of the inter-rater agreement scores indicated a moderate agreement between judges' rating of users' emotions. However, the agreement score of rating emotional valence was significantly higher. By comparing the three pairs of judges' agreement scores, the highest agreement score was between the external judges. In addition, comparing the agreement rating per emotion showed that rating confusion had the highest agreement among all other emotions. Compared with other emotions, delighted had the lowest

agreement scores. These results are consistent with the previous scientific literature [8] indicating that when using spontaneous facial expressions, delighted is the least predictable, and confusion is the most predictable emotion, compared with other affective states.

5.3 Emotions and Typing Features

The correlation analysis indicated that interaction features including number of attempts, length of response, session duration, and pause rate were moderately correlated with emotional change. The response verbosity (response quality and correctness) was also moderately correlated with emotional change. The results demonstrated weak correlation between users' affective state and typing speed, key latency, and key duration, and demonstrated moderate correlation with deletion rate, use of capitalization rate, and use of punctuation rate.

Our analysis of the correlation between typing features and the presence and absence of each emotion indicated that being delighted was associated with more careful writing, including use of punctuation marks, capitalization, deletion, and long correct answers. Being delighted was also associated with short pause rate and few tries. Confusion usually occurred after several interactions and was associated with careless writing behaviour. Confusion was also accompanied by long pause rate, fast typing, long key latency, and multiple attempts. Boredom was associated with long session duration, short pause rate, and use of unrelated keys.

While no typing features were associated with being neutral, frustration was associated with careless writing behaviours including less deletion rate, less use of punctuation, less use of capitalization, short responses, incorrect answers, and low quality answers. As well, frustration was associated with long pause rate, and multiple attempts.

5.4 Classification Assessment Summarization

The primary goal of this study was to explore the possibility of diagnosing users' emotions through their typing behaviours. The results of the classification confirmed the hypothesis that keystroke features provide a viable means to successfully determine emotional valence and moderately determine emotional states. The average classification accuracy of emotional valence using the standard classification methods were 69.17%, 63.19%, and 67.45% for the user-labelled, judge1, and judge2 dataset, accordingly. Using the same methods

to classify emotions yielded an average accuracy of 44.59%, 34.86%, and 43.59% for the user-labelled, judge1, and judge2 dataset, respectively.

Among all classifiers, the ANN classifier generated the most accurate results with an accuracy of 82.82%, 72.02%, and 77.2% on classifying emotional valence. On the other hand, the classification accuracy using ANN classifier on determining emotional states was 53.59% for user-labelled dataset, 45.6% for judge1 dataset, and 53.89% for judge2 dataset.

5.5 Hierarchical Classification and Affect Recognition Approaches

The overall classification accuracy of the hierarchical classifier was 59.37%, 49.74%, and 56.48% for the user-labelled, judge1, and judge2 dataset, respectively. Even though the hierarchical classification yielded a slightly better classification than the standard ANN classifier, the hierarchical classification accuracy outperformed the previous dialogue-based emotions classification method. The best classification accuracy yielded by the hierarchical classifier was 59.37% which is significantly better than the classification accuracy of similar approaches that used dialogue features to diagnose emotions. For example, D'Mello et al. [8] had a best classification accuracy of 42.4%.

The classification accuracy of the hierarchical classifier was comparable with the accuracy obtained from audio-based methods that used acoustic and lexical features to diagnose emotions. In general, the audio-based classifier that classified four to five emotions produced a classification accuracy range from 42.3%-78% using a variety of machine learning methods, including LDA, QDA, SVM, and fuzzy logic. For example, Kwon et al.'s [18] [4] classifier differentiated between five emotions using QDA, SVM, and LDA with an accuracy of 42.3%. Steidl et al.'s approach [35] differentiated among four emotions with an accuracy of 60%.

Comparing our approach with vision-based methods indicated moderate classification ability of users' emotions solely through typing features compared with vision-based approaches that use facial features and head gestures to determine emotions. Vision-based classification accuracy ranged from 39.58% to 84% e.g., Lee et al.[4] and Wang et al.[39].

Overall, a comparison of the costs and benefits of each method indicated that using typing features in the dependent context of a dialogue tutoring system sufficiently determined users' affective state without additional tools or computational expenses, in comparison with multimodal-based, vision-based or audio-based methods. Nevertheless, using typ-

ing features alone to diagnose users' emotions in an independent context requires further investigation.

5.6 Dimensionality Reduction Results

The result of applying principal component analysis on the three raw datasets indicated that three features accounted for 95% of the variance of judges' dataset and 96% of the user-labelled dataset. However, applying PCA on the three normalized datasets showed that eight features accounted for 77% of the variance in the judges' dataset and 80% of the variance in the user-labelled dataset. The classification of the raw data was slightly better than the dimensionally reduced data on all three datasets due to the independence among the features.

5.7 Normalization Results

Taking individual differences into account by normalizing the data per participants slightly improved the accuracy of k-nearest neighbours and discriminant analysis classification methods. Conversely, using the raw data improved classification results for most of the classification methods. On the basis of these results, using individual differences did not significantly improve the classification accuracy of determining emotions. However, this still seems a worthy direction for further study.

5.8 Limitations

This section discusses the lessons learned during this research process including technical limitations and theoretical challenges. Automated affect recognition with high accuracy is challenging due to lack of understanding of the nature of emotions. One factor affecting the emotional complexity is that several emotions have similar characteristics and may overlap. In the user emotion-labelling procedures, 33% of the vectors were associated with multi-emotions belonging to the same group (positive and negative). Most participants tended to experience overlapping emotions such as frustration, boredom, and confusion. These results suggest using dimensional descriptions to diagnose and report emotional states rather than categorical description of human affect.

One of the technical limitations during the self-emotion reporting procedure occurred when participants rated their emotions using a short survey asking whether or not they experienced each emotion. Using this method enabled subjects to choose multiple emotions sometimes belonging to different valences, which yielded 8% of the user-labelled data points associated with heterogeneous labels.

The results of the inter-rater scores suggested that a description for each emotional state was needed to reduce labelling discrepancy that completely depended upon raters' understanding of each emotional state. For example, some participants only reported positive emotions throughout the tutoring session, while others reported frustration after the first few questions. The divergence of the labelling manner could be attributed to the participants' varying cultural and experience backgrounds.

In addition, an uneven number of observations in each class resulted in poor classification in all of these classes. The user-labelled dataset consisted of 77 vectors associated with boredom and frustration, compared with 164, 80, and 146 data points associated with delight, neutral, and confusion. Judge1 dataset consisted of only 63 vectors associated with frustration, compared with 90, 162, 142, and 124 associated with delight, neutral, confusion, and boredom, respectively. Judge2 dataset also consisted of only 27 vectors associated with delight, and 164, 135, 168, and 87 data points associated with neutral, confusion, boredom, and frustrated, in turn.

5.9 Future Work

This section presents potential further study including additional improvements and other possible changes to be implemented on the existing dataset.This research investigated all possible keystroke features considered in keystroke dynamics, and affect detection studies [8][42][38]. However, current data presents a few more possible features that could be investigated. One possible modification would be that instead of calculating average typing speed, key latency, and key duration for each depressed key per answer, it could be worthwhile investigating the use of computed typing speed, key latency, and key duration for common English digraphs and trigraphs in sentences. The method of using typing speed, key latency, and key duration for the most common user digraphs or trigraphs to detect instability in user typing behaviour has been successful in a great deal of keystrokes dynamics research[12][38][37] .

Further studies could also be implemented on the existing datasets without needing to redesign the experiment. In this research a variety of machine learning techniques

were evaluated, but many other methods are also worthwhile investigating as they have produced high accuracies in affect and stress detection. These methods include Support Vector Machines (SVM), Hidden Markov Models (HMM), and AdaBoost. It could also be valuable to investigate the classification accuracy of diagnosing emotional states and emotional valence using fuzzy logic. These classifiers could be built based on the results of correlation analysis or Neuro-Fuzzy [2] that drives fuzzy rules from trained neural networks.

A possible improvement of the experimental design could include a weight for the user confidence rating. During the subjective ratings, the subjects provided an average of 30 feature vectors associated with the emotions categories. This means that subjects rated their emotions 30 times (on average) in half an hour which might have led to careless and inaccurate ratings. Thus, asking participants to evaluate their confidence in their ratings could provide more accurate labelling.

Chapter 6

Conclusion

In recent years, there has been increased interest in modeling human emotions to build intelligent computer systems able to recognize and adapt to users' emotions. Affect recognition studies have used multiple approaches to measure users' affective states such as facial expression, body position, heart rate, respiration rate, and pupillary size. Using typing features to detect users' affective states has many advantages over affective recognition methods requiring intrusive and expensive lab-based tools that are not feasible for everyday use.

Keystroke dynamics has been studied extensively in the field of computer security and moderately in affect and stress detection. Approaches comparing the reference model with user-provided typing to detect instability has been successful in the computer security field to detect unauthorized users. Similarly, using typing features to detect cognitive and physical stress was in line with affective recognition methods. However, past affect recognition approaches that used dialogue features only yielded a moderate to low accuracy compared with other affective computing approaches .

This thesis investigated the effectiveness of using timing and typing features to diagnose and model users' emotions in a dependent context. We focused on affective states that relate to learning which were spontaneously induced through a dialogue-based tutoring system that teaches computer-related topics in question-and-answer format. The tutoring system consisted of a user model, feedback generator, domain model, and diagnosis model. The system also computed and extracted the timing, typing, and response features. A field study was conducted to collect interaction data associated with subject-ratings and two external judges' ratings.

Twenty participants took part in the study, generating 544 data points in total for the

user-labelled dataset and 581 data points for the judges' datasets associated with emotions state and emotional valence rating. Six classification methods were evaluated on detection of users' emotional state and emotional valence. The emotional states were delighted, neutral, confused, bored, and frustrated, while emotional valence determined positive or negative emotions.

Despite the moderate to low correlations obtained from computing the correlation coefficient between typing features and users' emotions, the standard classification methods were able to successfully determine users' emotional valence on average classification accuracy of 69.17%, 63.19%, and 67.45% for the user-labelled, judge1, and judge2 datasets, in turn, and moderately classify emotions with an accuracy of 44.59%, 34.86%, and 43.59% for the user-labelled, judge1, and judge2 dataset, accordingly.

Among all the classifications methods, ANN classifier yielded the best classification compared to LDA, QDA, Naive Bayes, decision trees, and k-nearest neighbour with an accuracy of 82.82%, 72.02%, and 77.2% for diagnosing emotional valence intended for the user-labelled, judge1, and judge2 datasets, respectively. The classification accuracy for diagnosing emotional states using ANN were 53.59%, 45.6%, and 53.89% for the user-labelled, judge1, and judge2 datasets, respectively, which was improved from the standard classification methods.

All of the classification methods were implemented on the raw, normalized, and dimensionally reduced datasets. Using the dimensionally reduced data yielded a slightly lower accuracy, compared to the raw data attributed to the dependency between the features, where most of them accounted for some of the variation in the dataset. Similarly, using the normalized data according to user differences did not improve the accuracy of the classification except for linear discrimination analysis and k-nearest neighbour, where the classification accuracy for the normalized data was higher than the raw data.

To improve the classification accuracy of the neural network classifier, a hierarchical classification was implemented to first classify users' emotional valence into positive or negative emotions, next classify positive emotions into delighted or neutral, and finally classify negative emotions into confused, bored, or frustrated. The hierarchical classification generated an overall classification accuracy of 59.37%, 49.74%, and 56.4% on the raw data for the user-labelled, judge1, and judge2 datasets. While the hierarchical classification produced moderately higher accuracy than the standard neural network classification, it yielded a significantly higher accuracy in comparison with previous approaches that diagnosed users' emotions through their typing behaviour [8].

By using only timing, response, and typing features based on two values of key code and timestamp we were able to recognise when users experienced positive or negative emotions

with an average accuracy of 77.35% and determine which affective state users experienced with an accuracy of 55.20%. These results are in line with the accuracy obtained using audio-based features, but were lower than results obtained using vision-based features.

APPENDICES

Appendix A

Feedback letter and Recruitment scripts

Verbal and email script for student recruitment

Dear student,

We are looking for undergraduate and graduate students to partake in a research study that investigates the effectiveness of using keystroke features and the user's previous emotions that are detected by the system to assess his/her emotions during their ongoing interaction with an intelligent tutoring system.

In this study you will be asked to use CompTutor application, which is a dialogue-based intelligent tutoring system that teaches computer literacy through conversation. You will be first asked to fill out the user profile. Then, you will answer introductory level questions that related to computer hardware, software or information technology. All the questions are in open-ended form which allows you to type your answer in a text box. The session will be video recorded and the interaction data will be tracked. This study, in its entirety will take approximately half an hour of your time and you will receive $5 gift card as remuneration for your participation in the study.

If you are interested in participating please contact us by emailing Areej Alhothali at aalhotha@uwaterloo.ca . Any data pertaining to you as an individual participant will be kept confidential.

This study is supervised by Chrysanne DiMarco (School of Computer Science). The study is being conducted for Areej Alhothali's M.Math at School of Computer Science. This study was reviewed by, and received ethics clearance through, the Office of Research Ethics at the University of Waterloo.

Study Feedback

Date

Dear Participant,

We would like to take this opportunity to thank you for your participation in the research study. We want to specifically acknowledge your time and commitment to the study. It would not be possible to conduct this research without your participation.

Your participation has played a significant role in our research study the results of which will empower e-learning systems and intelligent tutoring systems with the ability to diagnose and respond according to the user emotions. Through this study we will investigate the effectiveness of using keystroke features and the user's previous emotions that are detected by the system to assess user's affective state during their ongoing interaction with intelligent tutoring system. This work may contribute to the body of knowledge in the area of affective computing and intelligent tutoring system.

All hard copies of consent forms and surveys will be stored under lock and key in the researcher's office. This will ensure that once collected, data with personal identifiers are securely stored in a locked area, and are accessible only to the research team.

An executive summary including the aggregated results of the study will be made available sometime in Jul 2011. We will send you a copy of this report via email.

This project has been reviewed by, and received ethics clearance through, the Office of Research Ethics. In the event you have any comments or concerns resulting from your participation in this study, please contact Dr. Susan Sykes at 519-888-4567, Ext. 36005.

Researcher Contact Information:

Areej Alhothali, Department of Computer Science
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
Tel: 519-888-4567 ext. 36657
aalhotha@uwaterloo.ca

David R. Cheriton School of Computer Science
University of Waterloo

PARTICIPANTS NEEDED FOR RESEARCH IN Affective computing in Computer science

We are looking for volunteers to take part in a study of
Modeling User affect from Interaction Events.

As a participant in this study, you would be asked to: *use an intelligent tutoring system (ITS) that teaches an introductory level of computer science. You will use the system for half-hour while the system is monitoring your ongoing interaction with ITS.*

Your participation would involve *1* session,
which is approximately *30* minutes.

In appreciation for your time, you will receive
$5 gift card

For more information about this study, or to volunteer for this study, please contact:
Areej Alhothali
David R. Cheriton School of Computer Science
519-888-4567 Ext. 34674 or
Email: aalhotha@uwaterloo.ca

This study has been reviewed by, and received ethics clearance through, the Office of Research Ethics, University of Waterloo.

Appendix B

Dialogue Scripts

	Level	Questions	Answers
Foundation of Information technology	Beginner	1. In computer network, what is the term Bluetooth referring to?	Bluetooth is a short-range wireless protocol for exchanging data.
		2. How confidential information should be sent using an unsecured network?	In an unsecured network, Information should be sent in encrypted format.
		3. Why it is important to update antivirus software regularly?	Updating anti-virus software is important to protect computers from all known viruses.
		4. What does the term Information and Communication Technology (ICT) mean?	ITC term covers any product that will store, retrieve, manipulate, transmit or receive information electronically in a digital form.
		5. How can you protect your computer against computer virus infection?	To protect your computer against viruses you should keep your antivirus software up dated.
		6. Describe how to safeguard your online identity.	Use a complex password for your e-mail and your computer .Use anti-spyware and anti-virus software .Don't give your personal information to strangers.
		7. What is the difference between hard and soft copy?	Soft-copy term refers to a document in digital format and hard-copy term refers to a printed document.
	Intermediate	8. In computer networks, what is the purpose of using a firewall?	A firewall is a computer system that prevents unauthorized access while permitting authorized communications.
		9. What are the differences between Intranet and Internet?	Intranet is a network that is not available (private network) to the world outside of the Intranet network. Internet is open to the public, to everyone who has an IP address.
		10. In communication, what does Wi-Fi refer to?	Wi-fi is a short for Wireless fidelity, a networking technology that allows devices to communicate without wires.
		11. What does the term computer hacking mean?	Computer hacking is the practice of modifying computer hardware and software to accomplish a goal outside of the creator's original purpose. For example ,people who try to break into protected or unprotected networks
		12. In computer security, what does the term malware refer to?	Malware is short for Malicious software, software designed to secretly access a computer system without the owner's informed consent.
		13. In computer security, what are the different types of malware?	Malware includes computer viruses, worms, Trojan horses, spyware.
		14. What does the term corrupted file refer to?	Corrupted files are computer files that suddenly become inoperable or unusable.
	Advanced	15. What are the differences between Intranet and Extranet?	An Intranet website is only available to the local network and not available to the world outside of it. However, an Extranet is actually an Intranet that is partially accessible to authorized outsiders.
		16. What does the term online piracy refer to?	Online piracy is a term used to describe the illegal copying of copyrighted materials from the Internet.
		17. What is an end-user licence agreement?	A software licence agreement is a contract between the licensor (manufacturer or author) and purchaser of the right to use software.
		18. What are the differences between e-commerce and e-business?	E-commerce refers to buying and selling using the Internet and e-business is the transformation of business processes through the use of Internet technologies.
		19. What does the term computer fraud refer to?	Computer fraud is the use of information technology to commit fraud
		20. What does the term data warehousing refer to?	Data warehousing is combining data from multiple and usually varied sources into one comprehensive and easily manipulated database.
		21. In computer security, what does the term data encryption refer to?	Encryption refers to algorithmic schemes that encode plain text into non-readable form or hypertext, providing privacy.

	Level	Questions	Answers
Computer Hardware	Beginner	1. What does the term computer hardware mean?	Computer hardware refers to the physical components of a computer.
		2. What does RAM stand for?	Random Access Memory
		3. What does ROM stand for?	Read Only Memory
		4. What does LAN stand for?	Local Area Network a relatively small network
		5. What does WAN stand for?	Wide Area Network
		6. What does computer networking refer to?	Computer Networking is the practice of linking computing devices (computers) together with hardware and software that supports data communications across these devices.
		7. What are the components of a central processing unit (CPU)?	The primary components of a computer CPU are the arithmetic logic unit (ALU), the control unit and the registers.
	Intermediate	8. What is RAM used for?	Random access memory used by the system to store data for processing by a computer's central processing unit (CPU).
		9. What is ROM used for?	ROM is memory containing hardwired instructions that the computer uses when it boots up, before the system software loads.
		10. What is BIOS program?	BIOS is a low level program used by your system to interface to computer devices such as your video card, keyboard, mouse, hard drive, and other devices.
		11. In the central processing unit CPU, What is Arithmetic/Logic Unit (ALU) used for?	The ALU executes the computer's commands by doing arithmetic or the logical comparisons.
		12. What does the auxiliary storage refer to?	Auxiliary storage is used to provide additional, portable storage that is available to a processor only through its input/output channels.
		13. What are the most commonly used transition media in computer networks?	The commonly used transition media in computer networks are Twisted wire, coaxial cable, and microwave.
		14. In computer networks, what is a modem used for?	Modems change incoming analog signals to digital signals, and outgoing digital signals to analog signals which are used to connect two computers using telephone lines.
	Advanced	15. In computer hardware, what is a dual core process?	Dual core process is a CPU with two separate cores on the same die, each with its own cache.
		16. In a computer network, what does the term bandwidth refer to?	Bandwidth is often used as a synonym for data transfer rate, which is the amount of data that can be transfer from one point to another in a given time period.
		17. In a computer networks, what does the term router refer to?	A router is a device or software that determines the next network point to which a packet should be forwarded toward its destination.
		18. What is the Ping command used for?	The ping command is used to verify that a particular Internet address exists and can accept requests.
		19. What are the different types of auxiliary storages devices?	There are many different types of auxiliary storage devices such as Internal Hard Disk Drive, External Hard Disk Drive, Optical Drive (CD, DVD) and USB Flash Drive.
		20. What does the term network topology refer to?	Network topology refers to the layout pattern of interconnections of the various elements of a computer network
		21. What are the differences between gateway, bridge, and router?	Gateway connects networks of different types. Bridge connects networks of the same type. Router connects several networks.(such as)

	Level	Questions	Answers
Computer Software	Beginner	1. What is word processor software used for?	A word processor is a computer application used for the production of any sort of printable material (documents).
		2. What is a spreadsheet application used for?	A spreadsheet application features a calculation, graphing tools, pivot tables and a macro programming language.
		3. What does the term a read-only file refer to?	A read-only file cannot be modified but, can be saved with a new name.
		4. What is a computer virus?	A computer virus is a computer program that can copy itself and infect a computer.
		5. What is the routine to shut down a non-responding application?	To shut down non-responding application press Ctrl + Alt + Delete, select the application in the Task Manager window and click End Task.
		6. What is an operating system?	Operating system is a program that conducts the communication between the various pieces of hardware and the applications.
		7. What are the differences between an operating system and application software?	The operating system acts as a platform that run all other software on the computer system and manage the communication between computers software and hardware .However, application software used to perform specific task.
	Intermediate	8. What is a query command used for?	A query command is used to filter database records to show just the ones that meet certain criteria or to arrange them in a particular order.
		9. Describe the content of a database?	Database consists of a collection of tables that store particular sets of data.
		10. What are the differences between computers file and folder?	Computer files store data, while folders store files and other folders.
		11. What does data compression refer to?	Data compression is the process of encoding information using fewer bits. In other words it is the process of converting data from one format to another format that is physically smaller in size.
		12. What is an XML file and what it is used for?	An XML file is an Extensible Mark-up Language file which is a file designed to define the rules for encoding documents. A mark-up language is used to annotate text or add additional information.
		13. What does GUI refer to?	A Graphical User Interface (GUI) refers to the computer graphical interface which gives users access to the resources by pointing and clicking.
		14. What is an EXE file?	EXE is the common filename extension denoting an executable file.
	Advanced	15. What is a relational database?	A relational database contains tables which are linked together.
		16. What is the difference between data and information?	Data refers to the raw or unprocessed information .Information refers to the processed information.
		17. In Database, what are queries?	A request for information or command from a database.
		18. In database, what are wildcards?	Wildcards are a way to match any combination of words or symbols
		19. What are the differences between viruses and spywares?	A computer virus spreads software, usually malicious in nature, from computer to computer. While, Spyware or adware design to collect information about the computer user without appropriate notice and consent.
		20. What does the term multimedia refer to?	The term multimedia refers to files that contain both of sound and images with text and graphics.
		21. In windows system, what are the differences between hibernate or standby (sleep) mode?	In sleep mode the data will be save in the RAM and the power supply is withheld. While, hibernate mode save the data into physical hard disk and the power turned off.

	Level	Questions	Answers
Internet Technologies	Beginner	1. What does web log (blog) use for?	Web log is a website used for posting news, current events, and online journals concerning people's own experiences.
		2. What do social networking sites usually provide?	Social networking sites allow users to share ideas, activities, events, and interests within their individual networks.
		3. What are the benefits of online data storage?	Online data storage service provides remote backup for data, easy to share and access.
		4. What does E-learning refer to?	E-learning refers to any Internet based learning that uses technology to enable people to learn anytime and anywhere.
		5. In computing, what is software plug-in?	Software plug-in is a miniature programs that plug into a host program for additional functionality.
		6. What does the term search engine referring to?	Search engine is a program that searches documents in the World Wide Web through keywords given by the user
		7. In internet, what is pop-up ad?	Pop-ups are new web browser windows to display advertisements.
	Intermediate	8. What does URL stand for?	Uniform Resource Locator
		9. What does an IP address refer to?	IP is short for Internet protocol which is a unique number assigned to each computer on the Internet.
		10. In computer security, what does malware stand for?	Malware is short for Malicious software.
		11. What does the term Push Email refer to?	Push email refers to the process of pushing the electronic mail through to the client without waiting for polling.
		12. What does the term Hyperlink refer to?	A hyperlink is a graphic or a piece of text in an Internet document that can connect readers to another web-page
		13. What does the term bandwidth refer to?	Bandwidth is the data transfer rate or in other words the amount of data that can be carried from one point to another in a given time period.
		14. What are the differences between freeware and shareware software?	Freeware term refers to software available to everyone free of charge, while shareware means that the software will be free for a while, but payment will eventually be necessary.
	Advanced	15. What is a digital certificate?	An attachment to an electronic message used for security purposes. / Use to verify that a user sending a message is who he or she claims to be, and to provide the receiver with the means to encode a reply.
		16. What does e-mail spam refer to?	An e-mail spam refers to any junk e-mail or unsolicited bulk e-mail (UBE).
		17. What are computer cookies?	A computer cookie is a small text file which contains the user login information, which is placed on the client computer by a website.
		18. What is the different between Trojan and spyware?	A Trojan is a program that does not appear to be destructive when it secretly performs another action like opening backdoor for hackers. While, spyware is a program that spy on the user interaction with system.
		19. What is the difference between HTTP and FTP protocol?	HTTP protocol is used to view websites while FTP is used to access and transfer files.
		20. In computer network, what does the term proxy server refer to?	A proxy server is a server that acts as an intermediary for requests from clients seeking resources from other servers.
		21. What does the P2P network refer to?	A P2P network is a network of personal computers that communicate with one another by running proprietary P2P software.

References

[1] D. Ayman and A. D. Goldshine. Blood pressure determinations by patients with essential hypertension: I. the difference between clinic and home readings before treatment. *The American Journal of the Medical Sciences*, 200(4), 1940.

[2] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 1994.

[3] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 2010.

[4] R. Chellappa, C.L. Wilson, and S. Sirohey. Human and machine recognition of faces: a survey. *Proceedings of the IEEE*, 83(5), may 1995.

[5] C. Conati and H. Maclaren. Modeling user affect from causes and effects. *User Modeling, Adaptation, and Personalization*, 2009.

[6] M. A. Davis. Understanding the relationship between mood and creativity: A meta-analysis. *Organizational behavior and human decision processes*, 108(1), 2009.

[7] C. Derbaix and C. Pecheux. Mood and children proposition of a measurement scale. *Journal of Economic Psychology*, 20(5), 1999.

[8] S. D'Mello, R. W. Picard, and A. Graesser. Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, 2007.

[9] S. K. DMello, S. D. Craig, A. Witherspoon, B. Mcdaniel, and A. Graesser. Automatic detection of learners affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1), 2008.

[10] P. Ekman. Are there basic emotions. *Psychological review*, 99, 1992.

[11] P. Ekman, R. W. Levenson, and W. V. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 1983.

[12] R. S. Gaines, W. Lisowski, S. J. Press, and N. Shapiro. Authentication by keystroke timing: Some preliminary results. *Rand Report R-256-NSF. Rand Corporation*, 1980.

[13] M. Hoque, M. Yeasin, and M. Louwerse. Robust recognition of emotion from speech. In *Intelligent Virtual Agents*. Springer, 2006.

[14] J. S. R. Jang and C. T. Sun. Neuro-fuzzy modeling and control. *Proceedings of the IEEE*, 83(3), 1995.

[15] E. S. John. Teaching, typing, talking two case studies. *CALL, culture, and the language curriculum*, 1998.

[16] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8), 2007.

[17] B. Kort, R. Reilly, and R. Picard. An affective model of interplay between emotions and learning. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, 2001.

[18] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee. Emotion recognition by speech signals. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[19] M. R. Lepper and M. Woolverton. The wisdom of practice: Lessons learned from the study of highly effective tutors. *Improving academic achievement: Impact of psychological factors on education*, 2002.

[20] D. T. Lin. Facial expression classification using pca and hierarchical radial basis function network. *Journal of Information Science and Engineering*, 22(5), 2006.

[21] H. R. Lv, Z. L. Lin, W. J. Yin, and J. Dong. Emotion recognition based on pressure sensor keyboards. In *IEEE International Conference on Multimedia and Expo*. IEEE, 2008.

[22] G. Mandler. Thought, memory, and learning: Effects of emotional stress. *Handbook of stress: Theoretical and clinical aspects*, 1993.

[23] J. D. Mayer. How mood influences cognition. *Advances in cognitive science*, 1, 1986.

[24] G. J. McLachlan and J. Wiley. *Discriminant analysis and statistical pattern recognition*. Wiley Online Library, 1992.

[25] F. Monrose and A. Rubin. Authentication via keystroke dynamics. In *Proceedings of the 4th ACM Conference on Computer and Communications Security*. ACM, 1997.

[26] F. Monrose and A. D. Rubin. Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*, 16(4), 2000.

[27] K. Muldner, W. Burleson, and K. VanLehn. yes!: Using tutor and sensor data to predict moments of delight during instructional activities. *User Modeling, Adaptation, and Personalization*, 2010.

[28] M. Pantic, A. Pentland, A. Nijholt, and T. Huang. Human computing and machine understanding of human behavior: a survey. *Artifical Intelligence for Human Computing*, 2007.

[29] R. W. Picard. *Affective computing*. The MIT Press, 2000.

[30] R. W. Picard and S. B. Daily. Evaluating affective interactions: Alternatives to asking what users feel. In *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches*, 2005.

[31] R. Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1, 1980.

[32] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1), 2003.

[33] T. A. Salthouse. Effects of age and skill in typing. *Journal of Experimental Psychology: General*, 113(3), 1984.

[34] A. Sarrafzadeh, S. Alexander, F. Dadgostar, C. Fan, and A. Bigdeli. how do you know that i dont understand? a look at the future of intelligent tutoring systems. *Computers in Human Behavior*, 24(4), 2008.

[35] S. Steidl, M. Levit, A. Batliner, E. Nth, and H. Niemann. "of all things the measure is man: Automatic classification of emotions and inter-labeler consistency. *In Proceedings of the IEEE ICASSP*, 2005.

[36] G. A. Tsihrintzis, M. Virvou, E. Alepis, and I. O. Stathopoulou. Towards improving visual-facial emotion recognition through use of complementary keyboard-stroke pattern information. In *International Conference on Information Technology: New Generations(ITNG)*. IEEE, 2008.

[37] M. Villani, C. Tappert, G. Ngo, J. Simone, H. S. Fort, and S. H. Cha. Keystroke biometric recognition studies on long-text input under ideal and application-oriented conditions. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2006.

[38] L. M. Vizer, L. Zhou, and A. Sears. Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies*, 67(10), 2009.

[39] J. Wang, L. Yin, X. Wei, and Y. Sun. 3d facial expression recognition based on primitive surface feature distribution. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2, 2006.

[40] R. Westermann, K. Spies, G. Stahl, and F. W. Hesse. Relative effectiveness and validity of mood induction procedures: a metaanalysis. *European Journal of Social Psychology*, 26(4), 1996.

[41] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 1987.

[42] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation*, 13(1), 2004.

[43] P. Zimmermann, P. Gomez, B. Danuser, and S. Schr. Extending usability: putting affect into the user-experience. *Proceedings of NordiCHI06*, 2006.

[44] P. Zimmermann, S. Guttormsen, B. Danuser, and P. Gomez. Affective computing–a rationale for measuring mood with mouse and keyboard. *International Journal of Occupational Safety and Ergonomics*, 9(4), 2003.