Identifying Emotional States using Keystroke Dynamics

Clayton Epp, Michael Lippold, and Regan L. Mandryk
Department of Computer Science, University of Saskatchewan
176 Thorvaldson Bldg., 110 Science Place, Saskatoon, SK, S7N5C9, Canada
clayton.epp@usask.ca, mike.lippold@usask.ca, regan@cs.usask.ca

ABSTRACT
The ability to recognize emotions is an important part of building intelligent computers. Emotionally-aware systems would have a rich context from which to make appropriate decisions about how to interact with the user or adapt their system response. There are two main problems with current system approaches for identifying emotions that limit their applicability: they can be invasive and can require costly equipment. Our solution is to determine user emotion by analyzing the rhythm of their typing patterns on a standard keyboard. We conducted a field study where we collected participants' keystrokes and their emotional states via self-reports. From this data, we extracted keystroke features, and created classifiers for 15 emotional states. Our top results include 2-level classifiers for confidence, hesitance, nervousness, relaxation, sadness, and tiredness with accuracies ranging from 77 to 88%. In addition, we show promise for anger and excitement, with accuracies of 84%.

Author Keywords
Affective computing, keystroke dynamics, emotion sensing

ACM Classification Keywords
H5.2.h [User Interfaces]: Input devices and strategies.

General Terms
Human Factors.

INTRODUCTION

Despite progress in graphics capabilities and processing power, interactive applications still have considerable usability problems. One main reason for these problems is that applications do not understand or adapt to users' context, such as their location, expertise, or emotional state. As a result, applications often act inappropriately: they provide inappropriate feedback, interrupt the user at the wrong time, and increase frustration. To solve these problems, we must make advances in two key areas: first, a set of mechanisms for gathering and modeling user context; and second, a set of techniques for adapting user interfaces and system behavior based on contextual information.

User context includes information such as the user's location, situation, or expertise, but one often ignored type of context that could radically change our computer interactions is the user's emotional state. If a user of a safety-critical application was frustrated or distracted, it could dangerously affect their performance. If a user of an online tutoring system was frustrated or distracted, the system could adapt its presentation of learning materials to better suit that student's learning style. For users of ordinary computer applications, frustration or distraction may not be dangerous, but can lead to increased errors. Systems that detect and respond to a user's emotional state could improve user performance as well as satisfaction. In addition, emotionally intelligent systems could also aid in computer-mediated communication by incorporating the user's emotional state into messages and by allowing users to naturally express emotional content to others.

Many approaches for detecting user emotions have been investigated, including voice intonation analysis, facial expression analysis, physiological sensors attached to the skin, and thermal imaging of the face. Although these explorations have seen varying rates of success, they still exhibit one or both of two main problems preventing wide-scale use: they can be intrusive to the user, and can require specialized equipment that is expensive and not found in typical home or office environments.

Our solution is to detect users' emotional states through their typing rhythms on the common computer keyboard. Called keystroke dynamics, this is an approach from user authentication research that shows promise for emotion detection in human-computer interaction (HCI). Identifying emotional state through keystroke dynamics addresses the problems of previous methods by using standard equipment that it is also non-intrusive to the user.

To investigate the efficacy of keystroke dynamics for determining emotional state, we conducted a field study that gathered keystrokes as users performed their daily computer tasks. Using an experience-sampling approach, users labeled the data with their level of agreement with 15 emotional states and provided additional keystrokes by typing fixed pieces of text. Our approach allowed users' emotions to emerge naturally with minimal influence from our study, or through emotion elicitation techniques.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
Copyright 2011 ACM 978-1-4503-0267-8/11/05....$10.00.

From the raw keystroke data, we extracted a number of features derived mainly from key duration (dwell time) and key latency (flight time). We then created decision-tree classifiers for 15 emotional states, using the derived feature set. We successfully modeled six emotional states, including confidence, hesitance, nervousness, relaxation, sadness, and tiredness, with accuracies ranging from 77.4% to 87.8%. We also identify two emotional states (anger and excitement) that show potential for future work in this area.

RELATED WORK

Modeling affective state using typing rhythms draws from two fields: affective computing and keystroke dynamics.

Affective Computing

Affective computing refers to "computing that relates to, arises from, or deliberately influences emotions" [27]. We are interested in identifying a user's emotional state, so we must first consider how emotions are described, and what other approaches have been used to classify emotion. The terms affect and emotion are often used interchangeably; we will use *emotional state* to refer to the internal dynamics (cognitive and physiological) that are present during an emotional episode, and *emotional experience* as what an individual perceives of their emotional state [27].

Describing Emotions

Two main approaches have been used to describe emotions: categorical and dimensional. The categorical approach applies specific labels to different emotional states through language (e.g. sadness, fear, joy) [12]. The dimensional approach [28] uses two orthogonal axes called arousal and valence. *Arousal* is related to the energy of the feeling and is typically described in terms of low (e.g. sleepiness) to high (e.g. excitement) arousal. *Valence* describes the pleasure (positive) or displeasure (negative) of a feeling. Labels for different emotional states can be represented in this two-dimensional space. For example, anger would be a high-arousal, low-valence state.

Sensing Emotional State

Both the categorical and dimensional models of emotion have been used in prior approaches of identifying emotional state. Some approaches use features easily discernable by other humans, such as facial expressions, gestures, vocal intonation, and language [27]. For example, face-tracking software is used to analyze facial expressions gathered from webcam images to infer users' affective states [9,26]. This approach has been extended to use thermal imaging to identify changes in blood flow patterns of the face that are synonymous with different facial expressions [22].

Other approaches use features that are less discernable to other humans, but can be measured by specialized equipment. For example, significant research has been conducted on measuring physiological changes that occur in the body during emotional episodes using sensors such as galvanic skin response, electromyography of the face, and heart activity (see [14] for an overview). In HCI, researchers have used physiological sensors to measure the affective state of a user interacting with technology. Results have been produced by studying users playing video games [24], navigating web pages [32], using video conferencing software [33], and using mobile technology [7].

The above approaches have two main problems that prevent their widespread use: the sensing technology is obtrusive, and requires expensive specialized equipment. For example, EKG is measured using electrodes attached directly to the user's skin. In some cases, the area where the electrodes are placed needs to be shaved to prevent interference [30]. Although research is underway to integrate these sensors into interaction devices, they are currently intrusive and their mere presence may alter the user's emotional state. In [22], a thermal camera is used is measure blood flow to a user's face. Although unobtrusive, the equipment is specialized and not found in typical home or office settings. To eliminate the need for intrusive and costly equipment, we propose to determine affective state via typing rhythms.

Keystroke Dynamics

Keystroke dynamics is the study of the unique timing patterns in an individual's typing, and typically includes extracting keystroke timing features such as the duration of a key press and the time elapsed between key presses.

Much of the previous research in keystroke dynamics has been in authentication systems, with the premise that, just as with handwritten signatures, the way that an individual types can be unique enough to identify them [21]. The use of keystroke dynamics for user authentication has been an active area of research, producing many studies [3,10,21,25], patents [2], and systems [1], whereby users are authenticated by providing the correct user name, password, and typing rhythm (see [13] for an overview). Anecdotal evidence suggests that strong emotional states can interfere with authentication [25]; however, little is mentioned of this and it is unclear whether the timing variance associated with these emotional states is similar between individuals.

Most of the authentication systems [3,21,25] use fixed-text models – that is, they use the same static piece of text (entered during authentication) that the model was trained on. There have been fewer approaches [10,16,25] that use models based on free text (text that is not prescribed to the user), as they do not perform as well as fixed-text models [25]. The length of the required training text varies between different studies; some require a few words [3] or full pages of text [15], which can create better performing models [4].

Although fixed-text models generally perform better than free-text models, the potential applications of free-text models are desirable. Recent work has explored free-text models for use in continuous verification, where users are continually monitored to identify masqueraders at any time (not just during authentication), and have shown potential

given enough samples of sufficient length [16]. Free-text models have even been able to identify individuals typing in different languages [17] as long as the two languages have enough similar valid digraphs. Most free-text studies require users to enter any 'valid' text as sample text [16]; however, in [10] keystroke activity was monitored as a background process during normal computer use. This method had three benefits: the user was less disturbed by the collection method, the data was obtained unobtrusively, and it reduced the cognitive load on the user by avoiding situations where they must think of something to type.

Classification algorithms for the analysis of keystroke dynamics for user authentication include neural networks [5], distance measures [21,25], decision trees [29], and other statistical methods [3,10,25]. Due to the differences in data collection approaches and classification methods, a comparison of performance across studies is difficult [3].

Keystroke Dynamics & Affective Computing

There has been very little previous work applying keystroke dynamics to affective computing.

Zimmerman et al. [35] describe a method to correlate user interactions (keyboard and mouse) with affective state. Affective states were induced using films. Physiological sensors were used in conjunction with the Self-Assessment Manikin (SAM) [23], a method of subjectively expressing affective state. The authors found significant differences between the neutral state and other emotional states, but were unable to distinguish between the induced states.

Recent work by Vizer et al. [31] used keystroke timing features of free text in conjunction with linguistic features to identify cognitive and physical stress. They achieved correct classifications of 62.5% for physical stress and 75% for cognitive stress (for 2 classes), which they state is comparable to other affective computing solutions. They also state that their solutions should be tested with varying typing abilities and keyboards, with varying physical and cognitive abilities, and in real-world stressful situations.

METHODOLOGY

The two primary components of this work are the data collection process and the data processing required to create classifications of emotional state. The data collection process consisted of gathering and labeling users' keystroke data. The data processing consisted of extracting relevant keystroke features to build classifiers.

Experience Sampling of Emotional Keystroke Data

We wanted to gather keystroke data *in situ* – as participants performed their daily computer activities – but also needed to label each data point with the emotional state of the user. To accomplish both of these goals, we used an experience-sampling methodology (ESM), whereby we periodically collect user keystroke data and user responses to emotional state questionnaires. In ESM [19], participants are asked to record their experiences periodically in real-time during their daily activities. The purpose is to gather temporal data 'in the moment' rather than retrospectively, which avoids problems of incorrect reconstruction or forgetfulness of the user. The drawback is that researchers cannot control or balance the different states tested.

We chose an experience-sampling methodology for two reasons. First, we were interested in emotional data gathered in the real-world, rather than induced in a laboratory setting through emotion-elicitation methods [8]. Our results are intended for use in real-world systems, and gathering the data for modeling from naturally occurring emotions increases our ecological validity. Second, as this is a new affect sensing technique, we wanted to explore a wide range of emotional states, and emotional induction, in the laboratory, is limited to one or two emotional states.

In experience-sampling studies, users are typically outfitted with a signaling device that alerts them to complete a self-report on their current state and/or situation [19]. We developed custom software for our signaling device and data collection that participants installed on their personal computers for use while they performed their daily tasks.

Data Collection Software

The data collection software was written in C# and used a low-level windows hook to scan each keystroke as it was entered by the user. This program ran in the background, gathering keystrokes regardless of the application that was currently in focus. The only visible sign that the application was running was an icon in the desktop system tray.

Based on the user's level of computer activity, the program prompted the user throughout their day. At each prompt, the user was presented with their keystroke text from the previous 10 minutes to review, then with an emotional state questionnaire, and then with some fixed text to type. The user could opt out of data collection at any time during the prompt if they were too busy or did not want to share their keystroke data (e.g., contained sensitive information like a password). Initial keystroke data was called *free text* as it was not constrained or influenced by our study.

The emotional state questionnaire contained 15 5-point Likert scale questions regarding a user's current emotional state: *I am frustrated, I am focused, I am angry, I am happy, I feel overwhelmed, I feel confident, I feel hesitant, I feel stressed, I feel relaxed, I feel excited, I am distracted, I feel bored, I feel sad, I feel nervous, I feel tired.* For each statement, the user would: *strongly disagree, disagree, neither agree nor disagree, agree,* or *strongly agree*.

The user was then asked to enter a randomly selected piece of text from *Alice's Adventures in Wonderland* [6]; this was how we collected the *fixed text* from users. To prevent copying and pasting, the user was unable to select the fixed text. The user then had a chance to review what s/he just entered before submitting the data remotely to our data collection server. The data collection server allowed us to preserve participant anonymity and supported remote users.

Field Study

The field study was conducted from July to October, 2009 with users participating for an average of four weeks. Participants installed the software on the computer that they used the most. There were no restrictions on their activities during the study; they had the freedom to work unimpeded.

Demographics

Upon installation of the software, the participant was presented with a one-time demographic questionnaire. We originally had 26 users take part in the study; however, not all participants completed enough study questionnaires to be included in the analysis – some completed as few as 2 over the duration of the experiment. Unlike a laboratory experiment, the field study did not allow us to control the amount of participation. To determine an informed threshold for inclusion in the study, we considered how many responses were needed to ensure that users were familiar with the study questionnaires and were calibrating their responses appropriately. Participants who completed more than one questionnaire per day on average for at least half of the study duration were likely familiar enough with the questionnaires and the process to be considered to have finished the study. Thus, we removed users with fewer than 50 responses, leaving 12 participants. This process occurred prior to the analyses and all remaining results are based on this reduced dataset. We had 10 male and 2 female users aged 24-34 (mean=28.5, s.d.=2.7), who were university students (9), admin personnel (2), and technicians (1).

Overall, participants were proficient with word processing, email, and instant messaging applications, with usages of 3-7 hours a week (3), 1-2 hours a day (2), and more than 2 hours a day (7). Ten participants indicated they spent at least half of their time on the computer that was collecting data. Work computers accounted for 8 of the installations with the remaining 4 installations on home computers. Most participants used desktop computers (10); few used laptops (2). One of these installations was on a virtual machine. We only included English-speaking participants because we focused on English character sequences in the features.

Feature Extraction

Once the data was gathered, we needed to transform the raw keystroke data files into a feature set that could be used as input to models. We extracted three different categories of information from the data: keystroke/content features, emotional state classes, and additional data points.

Keystroke Features

The raw keystroke data consisted of key press and release events, unique codes for each key, and a timestamp of when the key event occurred. Extensive processing grouped these keys into graphs of 2 or 3 symbols. Our keystroke features were mainly derived from the timing of single keystrokes as well as digraphs (two-letter combinations) and trigraphs (three-letter combinations). Initially, we extracted all key-specific features (e.g., duration of 'ie' digraph); however, this caused a number of problems. The number of features grew to over 100,000, making training unrealistic and our data set very sparse (e.g. one user may use the digraph 'aa' often and others may not, leading to missing data points).

To narrow the scope, we used only aggregate features for this analysis and removed all key/graph specific features. For each feature, we extracted the mean and standard deviation because during a sample period, the user could enter the same sequence of keys more than once (e.g., entering 'th' twice during the 10-minute sampling period). Keystroke features used in our models are shown in Table 1 and the categories of features are described further.

Code	Description
2G_1D2D	The duration between 1st and 2nd down keys of the digraphs.
2G_1Dur	The duration of the 1st key of the digraphs.
2G_1KeyLat	Duration between 1st key up and next key down of the digraphs.
2G_2Dur	The duration of the 2nd key of the digraphs.
2G_Dur	The duration of the digraphs from 1st key down to last key up.
2G_NumEvents	The number of key events that were part of the graph.
3G_1D2D	The duration between 1st and 2nd down keys of the trigraphs.
3G_1Dur	The duration of the 1st key of the trigraphs.
3G_1KeyLat	Duration between 1st key up and next key down of trigraphs.
3G_2D2D	The duration between 2nd and 3rd down keys of the trigraphs.
3G_2Dur	The duration of the 2nd key of the trigraphs.
3G_2KeyLat	Duration between 2nd key up and next key down of trigraphs.
3G_3Dur	The duration of the third key of the trigraphs.
3G_Dur	The duration of the trigraphs from 1st key down to last key up.
3G_NumEvents	The number of key events that were part of the graph.

Table 1. Coded keystroke features with descriptions.

Keystroke Duration Features (dwell): Keystroke duration features have been used extensively in previous keystroke dynamics work [3,4,5,17,25]. For our analysis, duration features were included for both single key features as well as graph features (i.e., digraphs and trigraphs).

For single keys, the duration was the time elapsed between 'key press' to 'key release'. Each key of the digraphs were extracted separately (e.g. 2G_1Dur was the duration of the 1st key of all digraphs). Key duration features included 2G_1Dur, 2G_2Dur, 3G_1Dur, and 3G_2Dur in Table 1.0. For graph duration features (2G_Dur and 3G_Dur), duration was measured as the time elapsed from the first 'key press' event to the last key's 'key release' event. For example, for the digraph 'the' the 3G_Dur would be the time from the 't' press event to the 'e' release event.

Keystroke Latency Features (flight): Keystroke latency features have been used in previous keystroke dynamics studies in authentication [3,5,25]. Keystroke latency is the time elapsed from one key release event to the next key press event or the 'time between keys'. Unlike the duration features, latency always involves two keys so there are no single key latency features. Our keystroke latency features (2G_1KeyLat, 3G_1KeyLat, 3G_2KeyLat) were separated for each pair of keys found in the graph.

Other Keystroke Features: We included features that combined some aspects of the duration and latency features. This included the "key down to down" features (2G_1D2D,

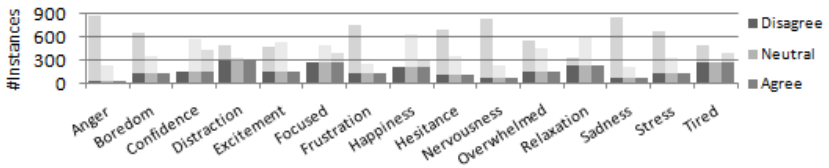

Figure 1. Distribution of responses (classes) before under-sampling (faded) and after.

3G_1D2D, 3G_2D2D), which contain the elapsed time from the first key down to the next key down event.

We calculated the number of events that were found in each digraph and trigraph (2G_NumEvents, 3G_NumEvents). Although most digraphs contain 4 events (first key press, first key release, second key press, second key release) and trigraphs contain 6, there are scenarios where the user could type more. For example, a user could press keys in quick succession and release them out of order (e.g. 1st key press, 2nd key press, 1st key release, 3rd key press, 2nd key release). This would result in 5 key events in a digraph. This may be indicative of a user's emotional state, and has not been used previously in keystroke dynamics, to our knowledge.

Keystroke Feature Overlap: Some of the keystroke features that we described overlap slightly; however, in the classification section, we describe how we reduced this set using supervised attribute selection.

Content Features

We included a few features based on the content extracted (text) from the free keystrokes, including separate features for the number of characters, numbers, punctuation marks, uppercase characters, and the number and percentage of 'special characters' (numbers, uppercase characters, punctuation marks). These features were used in only the free text models as we provided the fixed text to the user.

The number of mistakes (backspace + delete key) was calculated for both fixed and free text models. There are many different ways to correct mistakes (e.g., selection with the mouse and replacement with keystrokes). It was not possible to catch all of the possible correction scenarios as keystrokes were collected from different applications that we did not control. However, this feature does give a general idea of the number of mistakes being made.

Emotional State Classes

All of our features needed to be labeled with the user's emotional state for classification. We used discrete emotion categories to collect emotional state responses from users because these categories are close to the language commonly used to describe their emotional state.

As mentioned, we collected Likert Scale responses for 15 emotional states (e.g. "I feel stressed"). The available options that were presented for each statement (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree) became the target classes during classification. Many users avoided the 'extreme' categories of the 5-point scale, resulting in highly under-represented classes (class skew). We grouped the 'strongly agree' and 'strongly disagree' categories with the 'agree' and 'disagree' categories respectfully, resulting in 3 classes for each emotional state: agree, neutral, and disagree.

Additional Data Points

We extracted a number of additional data points to assist in our analysis, including the active process name for each collected keystroke, which would allow us to analyze our data set differentially depending on application.

Additional Data Processing

Our keystroke feature extraction did not take into account large pauses in typing, when a user might have switched to another mode of input (e.g. mouse) or have taken a break from the computer. To remove these pauses, we calculated outliers for all of the features that involved multiple keys (e.g. digraph latency). Outliers were removed by calculating the mean and standard deviation for all keystroke timing features for each participant and then removing the samples that were 12 standard deviations greater than the mean for that participant. A large number of standard deviations were used to only remove trials where very long pauses were present. We then recalculated all of the features with the filtered instance set. This resulted in the removal of 0.07% of the samples collected. This method of outlier removal was similar to the approach taken in [21]. All values were normalized for each user to facilitate the aggregate analysis.

CLASSIFICATION

Due to the large variations in the number of responses per user, we did not create user-specific models, but aggregated the data across participants for 1129 valid instances. Models were created using the C4.5 supervised machine learning algorithm as it is implemented in Weka [34].

Although there are many classification algorithms, we used decision trees as a simple low-cost solution to investigate this new approach to identifying emotion. Decision trees can be reduced into a set of rules and allow generalization (e.g. pruning). The chosen decision tree algorithm also handles missing values, which we have in our data set.

Supervised Attribute Selection

Although we identified 31 fixed text features and 37 free text features, we needed to first identify which of these features were important to keep and which had little predictive value. We used the correlation-based feature subset attribute selection method described by Hall in [18] and implemented in Weka [34] to select salient features in our set for each emotional state model separately.

Model Variations

We trained a number of model variations using different text types (i.e., free and fixed text), numbers of target classes, and adjustments to compensate for class skew.

In addition to the 3 class-level (agree, neutral, disagree) variations explained in the Feature Extraction section, we included variations that had 2 class-levels (agree, disagree), removing instances in the neutral category to determine if keystrokes could differentiate between two opposing states.

Responses were not distributed evenly across all levels of each emotional state (see Figure 1). For example, the responses to the phrase, "I am angry" were skewed to the strongly disagree and disagree categories. This is expected as users generally would not be very angry as often as not at all angry. Class skew such as this can lead to unreliable classification rates. To eliminate class skew, we used a method called under-sampling [11], which randomly removes instances from the majority classes to equal the class with the fewest instances, creating a uniform distribution (see Figure 1). We included models built using under-sampling and the original distribution. For each emotional state model that used under-sampling, we repeated the classification process 10 times and report the mean classification accuracy and variance.

Evaluation

We used 10-fold cross-validation from the stratified training results to evaluate our models, which is standard practice when the data set's size is limited [34].

To more easily describe our top results, we defined a hierarchy of evaluation categories: Bronze, Silver, Gold, and Platinum (see Table 2). Each category is based on the classification rates and the top three also incorporate false positive rate. To be considered for one of our categories, the sample-to-feature ratio had to be greater than 10. During the creation of the different model variations, instances were sometimes removed (e.g., instances of the majority classes were removed in under-sampling). In classifier design, it is generally accepted that there must be 10 times more instances (training samples) per class than the number of features [20]. Another criterion for inclusion was that the Kappa statistic had to be greater than 0.4. The Kappa statistic indicates how much the classification rate was a true reflection of the model or how much could be attributed to chance alone; Kappa values range from 0 (chance agreement) to 1 (perfect agreement) [34].

Type	Description
Bronze	Overall classification rate > 75%.
Silver	TP rates > 75%, FP rates < 25% for each class.
Gold	TP rates > 80%, FP rates < 20% for each class.
Platinum	TP rates > 85%, FP rates < 15% for each class.

Table 2. Evaluation categories used to describe the results.

RESULTS

We gathered both fixed text and free text data and created separate models for each of these data types. We are presenting the results of only the fixed text models as there were no free text models that made our top evaluation categories. For performance of the free text models, see [13]; in the discussion of this paper, we discuss potential improvement to the free text models. After removal of participants with fewer than 50 responses, the number of samples collected per participant ranged from 51 to 219 (mean=94.1, s.d.= 52.7).

Top Emotional State Models

Figure 2 shows the classification rates for our top-performing emotional state models. These top-performing models are all two-class models (agree and disagree), and perform appreciably better than chance (50%).

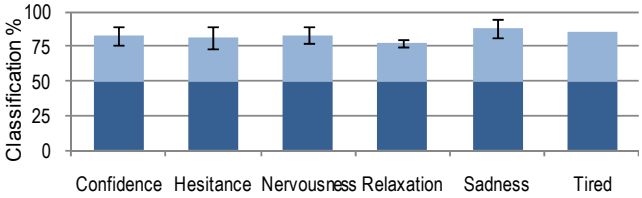

Figure 2. Classification rates for our top-performing models. Light bars show how our models improve upon random classification (chance). Error bars represent the variance in the classification rates after 10 random samplings.

We were very conservative in categorizing models as top-performing, to limit models to those with high classification rates, but also with little or no class skew, low false positive rates, high samples-to-features ratios, and high kappa values. We considered models that achieved our top 3 evaluation categories as our top emotional state models. Table 3 shows the results using 10-fold cross-validation. The Bronze category is not included because it was susceptible to class skew as it only looked at overall classification rates and not classification rates for individual class levels. Models in the Bronze category are presented in the next section as they show potential for future study.

Our top results were all based on 2 class levels (agree and disagree) and most of them used under-sampling. This means that these models were built using a reduced data set (reduced initially by removing the neutral category and further during under-sampling). Despite this reduction, we are still achieving satisfactory samples to features ratios (N:M) as seen in Table 3. N:M ranges from 11 to 49 with the most samples being used by the 'tired' model that used the original distribution of data (i.e., no under-sampling).

For 7 of the 8 top results in Table 3, under-sampling was used, and the correctly classified rate (CC) and Kappa statistic show averages from the 10 random-samplings models that were used during under-sampling. The variance columns in Table 3 (CC Var, Kap Var) indicate the variance from these 10 random samplings models. The 'tired' and 'relaxation' models have the lowest variance in both the classification rate and the Kappa statistic, and also have the highest sample-to-feature ratio, which may suggest that the variance can be lowered for the other emotional

State	CC	CC Var	Kap	Kap Var	M	N	N:M	Silv	Gold	Plat
Confidence	83.0	6.74	0.66	0.0027	8	286	18	X	X	
Hesitance	81.9	7.69	0.64	0.0031	9	204	11	X	X	
Nervousness	83.3	5.60	0.67	0.0022	5	152	15	X	X	
Relaxation	77.4	2.33	0.55	0.0009	8	442	28	X		
Sadness	87.8	6.52	0.76	0.0026	4	156	20	X	X	X
Tired	84.1	0.47	0.68	0.0002	9	758	42	X	X	
Tired*	85.1	0.00	0.70	0.0000	9	861	48	X	X	

Table 4. Top 3 evaluation category results. CC = correctly classified rate, Var = variance, Kap = Kappa statistic, M = #features used, N = #samples used, * = no under-sampling

state models given additional samples. Lower variance in the classification rates would suggest that the model results are truer indications of the predictive power of keystroke dynamics for emotional state detection.

When considering all of the factors (under-sampling, variance, sample-size) for the models in Table 3, the 'tired' model performs the best. The original distribution of responses for the 'tired' query was more balanced than the other emotional states (see Figure 1). This resulted in two 'tired' models reaching our top results, one using under-sampling and the other using the original distribution. These two models are quite similar, with similar numbers of instances, classification rates, and Kappa statistics.

Emotional State Models with Potential

As mentioned in the previous section, we considered emotional state models that made it to our Bronze category as emotional states with potential for future study, rather than as top-performing classifiers. This is because these models exhibited uneven distributions (class skew), which can lead to artificially-inflated classification rates. Each of the Bronze models in Table 4 are from unbalanced data.

Each emotional state from our top results in Table 3 also appeared in the Bronze category with an unbalanced version (no under-sampling), except for the tired state, which had an unbalanced version in Table 3. In addition, there were new representations from the 'excitement' and 'anger' emotional states. As Table 4 shows, classification rates were high; however, these models are also highly skewed. For example, the 2-state excitement model has an overall classification rate of 84.3%; however, when looking at the individual true/false positives for agree and disagree, we see that the 'disagree' class has a 92.7% classification rate whereas the 'agree' class has only 56.9%. Also, the 'disagree' class has 87.5% of the total number of samples. The same pattern exists for the anger emotional state.

It makes sense why such class skew exists in both of the anger and excitement states. These states have very high activation; it would be unusual for someone to be in a heightened state of anger or excitement for extended periods of time over the course of a normal workday. Because of the class skew for these two states, there were not enough instances remaining after under-sampling for these models to achieve our top-performing categories (see Figure 2). However, the high classification rates show that

State	CL	CC	Kap	D TP	D FP	N TP	N FP	A TP	A FP
Anger	3	83.9	0.53	92.0	38.4	64.6	9.2	0.0	0.0
Confidence	2	86.2	0.63	69.9	8.3	n/a	n/a	91.7	30.1
Excitement	2	84.3	0.53	92.7	43.3	n/a	n/a	56.7	7.3
Hesitance	2	92.3	0.67	94.9	25.5	n/a	n/a	74.5	5.1
Hesitance	3	76.3	0.55	86.9	26.9	63.3	10.4	48.0	6.5
Nervousness	2	93.0	0.52	96.9	48.7	n/a	n/a	51.3	3.1
Nervousness	3	81.6	0.51	95.2	40.7	46.5	5.4	38.2	3.4
Relaxation	2	78.9	0.56	82.4	26.2	n/a	n/a	73.8	17.6
Sadness	2	93.8	0.55	97.9	50.0	n/a	n/a	50.0	2.1
Sadness	3	82.9	0.55	93.8	33.3	56.0	6.2	35.9	3.9

Table 3. Emotional state models in the Bronze category. CL = class-level, CC = correctly classified rate, Var = variance, Kap = Kappa statistic, D = disagree, N = neutral, A = agree, TP = true-positive, FP = false-positive.

anger and excitement should not be ruled out as candidates for keystroke-based emotion modeling simply because of class skew in our particular data set. Future studies on these emotional states should consider using a laboratory-based emotion-elicitation method [8] rather than a field study.

Selected Features

Among the top emotional state models, the number of features were reduced from 31 (fixed text) to an average of 7.4 (s.d.=2.1). Using the correlation-based feature subset attribute selection method [18] allowed us to increase the sample-to-feature ratio with minimal loss to the classification rate. Table 5 lists the features that were used for each of the top classifiers (see Table 1 for a description of each feature). These features contain a fairly even number of both key latency and duration features with the 2G_1KeyLat_Mean and 2G_2Dur_Mean used in most of the models. Features that were not used in any of the models include the means of 3G_2Dur, 2G_Dur, and 3G_2KeyLat as well as the standard deviations of 2G_1D2D, 2G_1KeyLat, 2G_Dur, 3G_1D2D, 3G_1Dur, 3G_1KeyLat, 3G_2D2D, 3G_3Dur, 3G_Dur as well as our content features (NumMistakes).

Feature	C	H	N	R	S	T
2G_1D2D_Mean	X			X		X
2G_1Dur_Mean			X		X	X
2G_1Dur_Std						X
2G_1KeyLat_Mean	X	X	X	X	X	
2G_2Dur_Mean	X	X	X	X	X	X
2G_2Dur_Std		X		X		
2G_NumEvents_Mean	X			X		X
2G_NumEvents_Std	X	X			X	
3G_1D2D_Mean	X	X		X		X
3G_1Dur_Mean	X					
3G_1KeyLat_Mean		X	X			
3G_2D2D_Mean	X					X
3G_2Dur_Std		X				
3G_2KeyLat_Std				X		
3G_3Dur_Mean		X				
3G_Dur_Mean						X
3G_NumEvents_Mean			X			
3G_NumEvents_Std		X		X		X

Table 5. Features selected for each top emotional state model. C=Confidence, H=Hesitance, N=Nervousness, R=Relaxation, S=Sadness, T=Tired.

DISCUSSION

Our results show that keystroke dynamics can accurately classify at least two levels of seven emotional states (confidence, hesitance, nervousness, relaxation, sadness, and tired). In addition, we identified two other emotional states (anger, excitement) that have potential for keystroke-based classification. In this section, we discuss the benefits and drawbacks of using experience-sampling, the use of fixed text versus free text in keystroke-based modeling, aggregate versus individual analyses, the limitations of our data set, and opportunities for extensions to our work.

Experience-Sampling for Emotion Modeling

One of the goals of our research was to fill the gap in the related literature on the real-world applicability of affective computing solutions. Our choice to use keystroke dynamics was partly guided by this goal to create classifiers that are unobtrusive and inexpensive enough to be deployed in users' homes or workplaces. Experience-sampling allowed us to focus on the eventual application of our keystroke-based approach. Users' emotional states emerged naturally as compared to emotion-elicitation experiments [8].

Experience-sampling also provided us with the opportunity to perform exploratory research in this new field. Before our study, there was no guidance on which emotional states might be identifiable using keystrokes; experience-sampling allowed us to take a broad approach to the problem and narrow down which emotional states we should target with future work. This broad approach would be unfeasible using emotion elicitation because participants would have to be induced into each emotional state (neutral states as well) in individual experiments. In addition, mood induction does not necessarily work on all participants. Collecting data using experience-sampling and remote data submission allowed us to collect labeled emotional state data with minimal administration overhead.

However, experience-sampling introduced some drawbacks with implications for our analyses. Due to the uncontrolled nature of this methodology, we could not balance the distribution of classes for each emotional state, which led to class skew in the data. These uneven distributions limited our interpretation of the results for unbalanced models, and limited the number of instances available for our balanced models that used under-sampling.

Despite this disadvantage, we feel that experience-sampling provides a data-collection methodology that can be beneficial when studying a wide range of naturally-emerging emotions in new areas of research.

Free Text Models

Our free text models did not perform well enough to achieve our top evaluation categories and were not included in the results section. We believe that this was due to setting the user activity threshold too low. A low activity threshold caused the questionnaire to be presented to the user with fewer free text keystrokes than needed for analysis. Although the mean number of collected keystrokes were similar between the free (169) and fixed text (166), the standard deviation for the free text was quite high (302.8). A high standard deviation implies that some samples had very few free text keystrokes whereas the fixed text remained relatively consistent. Future studies should ensure that enough free text data is collected.

Aggregated versus Individual Participant Analyses

We created emotional state models for the entire data set across all participants to maximize the number of samples, especially in the under-sampled versions of the models. Creating models at the level of individual participants was not viable due to the number of collected samples with which to work. The mean number of samples per participant was 94.1, and this number would have been reduced when adjusting for class skew using under-sampling and in 2 class-level models. Having so few instances would have caused our sample to feature ratio to be too small and our results to be less accurate and reliable.

However, we do not know whether there are large individual differences in how keystroke dynamics change with emotional state. With enough samples per participant, personalized models could improve our classification rates. The success of keystroke dynamics for user authentication is based on the fact that each person's typing rhythm is unique enough to identify them. This suggests that individuals might have unique keystroke-level reactions to different emotional states. For example, when stressed, some individuals may type faster and other may pause in their typing more frequently. Accounting for these personal differences may allow us to build better performing models.

Limitations of our Data Set

We took care to ensure that we did not artificially inflate our classification rates by ignoring the two main limitations of our data set: its limited size and its uneven distribution.

Half of the participants in our study submitted fewer than 50 samples over the course of 4 weeks. For ethical reasons, we allowed participants to opt out of submitting data for each sample, but we tried to encourage participation through the use of data-based incentives (for each week that users submitted at least 15 samples, they were entered into a draw for a cash prize). After including only the active participants (more than 50 total samples), we had 1129 samples over all participants. This total was further reduced in the 2 class-level models though the elimination of the neutral category and in the under-sampled models by balancing the samples per class. In future work, we would prefer larger data sets with enough samples to perform individual-level models. Collecting user data for a longer time period and providing better incentives for active participation would increase our data set size.

The second problem with our data set was unequal distributions of responses across some of the emotional states. This is understandable as our feelings over the

course of the day are not evenly distributed. To address this problem of class skew, we used under-sampling, which reduced our data set size. In our field study software, we prompted the user at random times to fill out the experience-sampling questionnaire. We also allowed them to select the questionnaire explicitly if they wanted to submit their data. Coaching participants to use this explicit submit feature when they were experiencing low-frequency emotions like anger could help improve the distribution of responses. With the preliminary models provided by our work, we could deploy an adaptive version of the software that conducts modeling in the background and prompts the experience-sampling questionnaire when it detects that the user might be in a low-frequency emotional state.

Opportunities for Future Work

Our work is the first to classify a broad range of emotional states using typing rhythms. We have demonstrated the efficacy of this technique, which opens up opportunities for future research to refine, improve, and utilize this approach.

Because of our limited data set, we used all of the available samples in our models. Filtering the samples based on the application context could improve classification rates. For example, the keystroke dynamics gathered using word processing software are likely different than when writing code in an integrated development environment. Samples could be filtered based on the application that was active during each keystroke entered, which would allow for application-specific models. Although we did not filter samples based on the active application, our software did collect this information in anticipation of future analysis. In addition, we would like to investigate how models personalized to individual users change classification rates, as described in the section on aggregated analysis.

When selecting our features, we used only generalized features, such as the flight time between two keys in a digraph. We could also investigate specific features, such as the flight time between two keys in a specific digraph (e.g., 'th'), or a number of specific digraphs (e.g., the 20 most common digraphs in the English language). We calculated these features from our raw data [13], but due to our data set size, we focused on the general features.

Rather than use individual features in a model directly, feature aggregation through an approach like Principle Component Analysis (PCA) could be used. We investigated using PCA to select and aggregate features, but found that this process did not improve overall classification rates [13]. In addition, the resulting decision trees were difficult to interpret due to the feature aggregation. We could also use aggregation of the target classes to improve our results. Informal analysis of the 15 emotional states found many correlations. Combining the co-varying states could improve classification rates and the potential application.

Adding other interaction-based features could improve classification. Adding linguistic features as model attributes (similar to [31]) is an approach that could be applied to the free-text models (fixed text and the resulting linguistic features is, by definition, prescribed). Adding low-level features based on mouse kinematics might also improve classification. In addition, using a multi-modal approach could provide better generalization of the models to different computer-based tasks that require varying amounts of typing or aiming. Our software collected linguistic data and mouse kinematics, in anticipation of future analysis. We could also combine our approach with established methods (e.g., physiological models) for a more complete and robust model of user experience.

We used our own custom questionnaire to take a broad approach to collecting a variety of user states. Using a validated emotional state scale [8] would provide additional validity to the keystroke dynamics approach. Also, distinguishing between different emotions rather than levels of a single emotion would be beneficial.

Finally, although we used decision trees as our classification algorithm, there are other machine learning algorithms (e.g. support vector machines) that might provide better-performing classifiers.

CONCLUSION

The ability to recognize emotions is an important part of building intelligent computers. Systems that could extract the emotional aspects of a situation would have a rich context from which to make appropriate decisions about how to interact with the user or adapt their system response. There are two main problems with current approaches for identifying emotions that limit their applicability: they can be invasive and may require expensive equipment. We presented a solution that determines user emotion by analyzing the rhythm of users' typing patterns on a standard keyboard. To gather emotionally-labeled data, we conducted a field study where participants' keystrokes were collected and their emotional states were recorded via self report using an experience-sampling methodology. From this data, we extracted keystroke features, and reduced our feature set using correlation-based feature subset attribute selection. We created classifiers for 15 emotional states.

Our top results include 2-level classifiers for confidence, hesitance, nervousness, relaxation, sadness, and tiredness with accuracies ranging from 77.4 to 87.8%. In addition, our results show promise for anger and excitement, with accuracies of 84%. This work presents the first use of naturally-gathered typing rhythms to identify the emotional state of a computer user, and the first method of sensing a variety of emotional states unobtrusively and inexpensively. This work is important as it moves us closer to creating emotionally-aware computers that can be widely deployed.

ACKNOWLEDGEMENTS

Thanks to NSERC for funding and the University of Saskatchewan HCI lab for feedback.

REFERENCES

1. Admit One Security. *AdmitOneSecurity*. http://www.admitonesecurity.com.
2. Bender, S. and Postley, H. Key sequence rhythm recognition system and method. .
3. Bergadano, F., Gunetti, D., and Picardi, C. Identity verification through dynamic keystroke analysis. *Intell. Data Anal. 7*, 5 (2003), 469-496.
4. Bergadano, F., Gunetti, D., and Picardi, C. User authentication through keystroke dynamics. *ACM Trans. Inf. Syst. Secur. 5*, 4 (2002), 367-397.
5. Brown, M. and Rogers, S.J. User identification via keystroke characteristics of typed names using neural networks. *Int. J. Man-Mach. Stud. 39*, 6 (1993), 999-1014.
6. Carroll, L. *Alice's Adventures in Wonderland*. The Gutenberg Project, 2008.
7. Chen, D. and Vertegaal, R. Using mental load for managing interruptions in physiologically attentive user interfaces. *CHI '04 ext. abst. on Human fac. in comp. systems*, ACM (2004), 1513-1516.
8. Coan, J. and Allen, J. *Handbook of Emotion Elicitation and Assessment*. Oxford University Press, New York, USA, 2007.
9. De Silva, L. and Suen Chun, H. Real-time facial feature extraction and emotion recognition. *Infor., Comm., and Sig. Proc. 2003 and the 4th Pac. Rim Conf. on Multimedia*, (2003).
10. Dowland, P. and Furnell, S. A Long-term trial of keystroke profiling using digraph, trigraph, and keyword latencies. In *IFIP Intern. Fed. for Infor. Processing*. Springer Boston, 2004, 275-289.
11. Drummond, C. and Holte, R. C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. (2003).
12. Ekman, P. *Basic Emotions*. John Wiley & Sons, Ltd, 2005.
13. Epp, C. Identifying emotional states through keystroke dynamics. 2010. http://library2.usask.ca/theses/available/etd-08312010-131027/.
14. Fairclough, S. Fundamentals of physiological computing. *Inter. with Comp. 21*, 1-2 (2009), 133-145.
15. Gaines, R., Lisowski, W., Press, S., and Shapiro, N. Authentication by keystroke timing: some preliminary results. 1980.
16. Gunetti, D. and Picardi, C. Keystroke analysis of free text. *ACM Trans. Inf. Syst. Secur. 8*, 3 (2005), 312-347.
17. Gunetti, D., Picardi, C., and Ruffo, G. Keystroke analysis of different languages: a case study. In *Lecture Notes in Computer Science*. Springer Berline / Heidelberg, 2005, 133-144.
18. Hall, M. Correlation-based feature subset selection for machine learning. 1999.
19. Hektner, J., Schmidt, J., and Csikszentmihalyi, M. *Experience Sampling Method: Measuring the Quality of Everyday Life.* Sage Publications, Thousand Oaks, 2007.
20. Jain, A., Duin, R., and Mao, J. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 1 (2000), 4-37.
21. Joyce, R. and Gupta, G. Identity authentication based on keystroke latencies. *Commun. ACM 33*, 2 (1990), 168-176.
22. Khan, M.M., Ingleby, M., and Ward, R.D. Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations. *ACM Trans. Auton. Adapt. Syst. 1*, 1 (2006), 91-113.
23. Lang, P. Behavioral treatment and bio-behavioral assessment: computer applications. *Technology in mental health care delivery systems*, (1980), 119-137.
24. Mandryk, R.L. and Atkins, M.S. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *Int. J. Hum.-Comput. Stud. 65*, 4 (2007), 329-347.
25. Monrose, F. and Rubin, A.D. Keystroke dynamics as a biometric for authentication. *Future Gener. Comput. Syst. 16*, 4 (2000), 351-359.
26. Partala, T., Surakka, V., and Vanhala, T. Real-time estimation of emotional experiences from facial expressions. *Interact. Comput. 18*, 2 (2006), 208-226.
27. Picard, R.W. *Affective Computing*. MIT Press, Cambridge, 2007.
28. Russell, J. Core affect and the psychological construction of emotion. *Psychological Review 110*, 1 (2003), 145-172.
29. Sheng, Y., Phoha, V., and Rovnyak, S. A parallel decision tree-based method for user authentication based on keystroke patterns. *IEEE Transactions on Systems, Man, and Cybernetics 35*, 4 (2005), 826-833.
30. Stern, R.M., Ray, W.J., and Quigley, K.S. *Psychophysiological recording*. Oxford University Press, New York, 2001.
31. Vizer, L.M., Zhou, L., and Sears, A. Automated stress detection using keystroke and linguistic features: An exploratory study. *Int. J. Hum.-Comput. Stud. 67*, 10 (2009), 870-886.
32. Ward, R.D. and Marsden, P.H. Physiological responses to different web page designs. *Int. J. Hum.-Comput. Stud. 59*, 1-2 (2003), 199-212.
33. Wilson, G. and Sasse, M. Do users always know what's good for them? Utilizing physiological responses to assess media quality. (2000), 327-339.
34. Witten, I. and Frank, E. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2005.
35. Zimmermann, P., Guttormsen, S., Danuser, B., and Gomez, P. Affective computing - a rationale for measuring mood with mouse and keyboard. *Inter. J. of Occ. Saf. and Ergo. 9*, 4 (2003), 539-551.