IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 28, 2018, accepted July 2, 2018, date of current version August 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2854966

A Machine Learning Approach to Leverage Individual Keyboard and Mouse Interaction Behavior From Multiple Users in Real-World Learning Scenarios

SERGIO SALMERON-MAJADAS[1], RYAN S. BAKER[2], OLGA C. SANTOS[1], AND JESUS G. BOTICARIO[1]

[1]aDeNu Research Group, UNED, 28040 Madrid, Spain
[2]Graduate School of Education, University of Pennsylvania, Philadelphia, PA 19104, USA

Corresponding author: Sergio Salmeron-Majadas (sergio.salmeron@dia.uned.es)

This work was supported by the Spanish Ministry of Economy and Competitiveness through FPI grants under Grant BES-2012-054522 and Grant EEBB-I-16-11857 and projects under Project TIN2011-29221-C03-01 (MAMIPEC) and Project TIN2014-59641-C2-2-P (BIG-AFF).

ABSTRACT There is strong evidence that emotions influence the learning process. For this reason, we explore the relevance of individual and general mouse and keyboard interaction patterns in real-world settings while learners perform free text tasks. To this end, we have modeled users' mouse movements and keystroke dynamics with data mining techniques, building on previous related research and extending it in terms of some critical modeling issues that may have an impact on detection results. Inspired by practice in affective computing where physiological sensors are used, we argue for the creation of an interaction baseline model, as a reference point in the way how learners interact with the keyboard and mouse. To make the proposed affective model feasible, we have adopted a simplified 2-D self-labeling approach for labeling the users' affective state. Our approach to affect detection improves results when there is a small amount of data instances available and does not require additional affect-oriented tasks from the learners. Specifically, learners are only asked to self-reflect their emotional state after finishing the tasks and immediately selecting two values in the affect scale. The approach we have followed aims to distill two types of interaction patterns: 1) within-subject patterns (from a single participant) and 2) between-subject patterns (across all participants). Doing this, we aim to combine both the approaches as modeling factors, thus taking advantage of individual and general interaction patterns to predict affect.

INDEX TERMS Affective computing, data mining, keystrokes dynamics, learner modeling, learning analytics, MOKEETO, mouse movements.

I. INTRODUCTION

There is strong evidence that emotions have an important effect on learning outcomes [1]. For this reason, detecting learners' affect has been a pertinent research issue for decades [2]. This has been addressed in many different ways, covering diverse types of signals (e.g., electroencephalography, physiological responses, facial and body movements, voice features, verbal communication, keyboard and mouse interaction data) employed individually or in combination [3]. In most cases, emotions are inferred by classical machine learning/data mining approaches, which depend on data-driven approaches [4].

In this work we explore non-intrusive approaches for emotion detection, and thus investigate to what extent two traditional interaction sources, i.e., keyboard and mouse, which are common tools in many learning settings, can be further exploited. Although these two information sources have been previously used in affective state detection [5], [6], their attributes have not yet been systematically studied from a modeling perspective. We are particularly interested in affect detection within real-world learning scenarios, involving ordinary daily practices of learners where intrusive solutions are not as appropriate. In this specific case we are investigating the domain of learning English as

second language, which in many cases requires writing compositions in free text (i.e. the participants do not have to reproduce a given fixed text). Previous studies on related problems involved typing a fixed text several times [5], [7] or the induction of affect through stories or video clips [8]. Affect detection from free text data in realistic scenarios (i.e. participants performing their normal computer use) has resulted in lower accuracy results [5] compared to previous fixed text studies [5], [7]. It is very likely that users may use the mouse and keyboard to different degrees in realistic scenarios versus more artificial scenarios [7]. Other factors that may impact on these differences between settings is the variability in the number of interactions over time in the realistic scenarios (from where free text is collected) [5].

There is related evidence showing that affective state users' models built from keyboard and mouse and which focus on a user's individual interaction patterns tend to be more accurate [9], albeit in less natural settings. This result suggests that affect detection methods may be more accurate if they characterize the individual features of a person [2]. Modeling an individual person's behavior from keyboard and mouse has several challenges, such as the lack of large interaction data sets (especially in learning contexts) necessary to obtain an accurate model of the learner [9]. Individuals might have unique keystroke-level reactions to different emotional states [5]. Therefore, although searching for general affect interaction patterns has been relatively successful [6], getting personal patterns (which is suggested to improve the accuracy) is more challenging. This is because the accuracy of the methods used is strongly related to the size of the available samples [10], and thus, it might not be straight forward to obtain the expected improvement in the accuracy.

Given the aforementioned advantages of using individual interaction patterns and following a common practice in experiments using physiological data sources [11], the novelty of this work resides in exploring the effects of using an interaction baseline model for the keyboard and mouse interaction data sources. This interaction baseline model is based on the users' interactions with the keyboard and mouse and analyzed at the outset, where it is assumed that the students are not yet affectively altered by the learning process. Particularly, in the current paper, we propose the following hypothesis:

H: "Creating an interaction baseline model describing how each participant generally interacts with the keyboard and mouse improves affective detection results".

In our proposal, this baseline model represents how the user interacts with the keyboard and mouse, and is obtained from an initial task, called the calibration task. Using this modeling approach, the purpose is to take advantage of individual user features (in terms of interaction dynamics), while also leveraging features that are general across participants [5], [6]. This enables to take advantage of what can be learned from modeling each individual's interaction dynamics [9], while addressing the shortage of data on individual students [10].

The main goal of this work is to predict the affective state of learners, and to do this we focus on obtaining a readily available model for quickly detecting affective states, in which students label their own affective state using a well-known and widely used psychologically validated scale that represents the intensity of their affect reactions in two different dimensions, namely valence and arousal [9], [12].

Keyboard and mouse have been extensively employed in affect detection studies using data mining techniques. However, comparing different case studies is difficult because i) different emotional representations are used, ii) different feature subsets are considered in each study, and iii) different methodological decisions are taken in consideration when processing the datasets and generating the models [9]. To deal with this and further exploit the data mining process involved we are taking various methodological variables into account. To address the lack of individual user data in educational contexts [13] and the many features to describe users' interactions, in this work we have evaluated the efficacy of different preprocessing and reduction techniques to address this high dimensionality problem.

All these issues and other related topics are discussed in this paper through a real-world case study based on detecting the affective state of the user from keyboard and mouse interactions. In particular, this paper investigates a set of key modeling issues that are not commonly considered in prior work. The contribution of this paper is to explore the following issues: 1) using a baseline model of the user's individual keystroke and mouse dynamics, 2) reducing the high number of mouse and keyboard features and class balancing the data instances by applying data preprocessing techniques as methodological variables, and 3) adopting a simplified two-dimensional approach for labeling the user affective state. To account for these issues the rest of this paper is organized as follows. First we summarize related work, pointing out the main issues involved considering both traditional interaction devices (i.e., keyboard and mouse). Then we present the experiment that was implemented, including the methodology proposed for affective state detection, along with the tools, features and metrics used. Subsequently we report on the results of the experiment, which are analyzed in the context of related research in the discussion section. Finally, we present conclusions and outline future work.

II. RELATED WORKS

The literature review shows that over the last few years, affective state detection increasingly began to incorporate sensors that measure physiological signals [14]. However, in this work we follow a sensor-free approach (i.e. using the basic devices in a desktop setup as the only information sources and without requiring any additional hardware with monitoring purposes,) and aim to study the potential of keyboard and mouse interaction devices as information sources. For this reason, we have focused this review on studies that address specific keyboard and/or mouse modeling issues used for affective state detection. More specifically,

we are going to highlight how these studies approach the following methodological issues that we are going to discuss in our study (see subsection II.E for more details): i) interaction modeling features used for each device (and how those features are generated), ii) temporal applicability of the model (time window length used to collect and process data to generate the affective model), iii) affective labeling approach (and how it is collected), iv) emotional elicitation method, v) task proposed for data collection (experimental context aiming to elicit emotions), and vi) the number of data instances used. Notice that each study analyzed here provides different levels of information. Thus, the current paper only describes each study based on the information provided. Nonetheless, Table 1 and Table 2 are included after the analysis of related works to show similarities and differences between the features used across studies.

The review of the literature is reported as follows. The next subsection describes related works that propose keyboard as the main data source for affective state detection. The following subsection includes works which use the mouse as the main data source. After that, works combining both mouse and keyboard features are discussed. Once the different works have been identified, a subsection is included to compile and structure the different features proposed in the literature. The last subsection discusses the methodological issues identified in the works that have been reviewed.

A. KEYSTROKE ANALYSIS

The keyboard is one of the most common data sources employed in sensor-free approaches to affective state detection. Keyboard interactions have been traditionally used for biometric purposes (aiming to identify users from their unique way of typing), but there is evidence that models based on keystrokes dynamics generated for biometric purposes exhibit instabilities due to transient factors such as emotion, stress, drowsiness, etc. [15]. These instabilities in the user model, caused by emotions, have triggered some studies analyzing keystroke analytics for affective detection purposes [16], [17].

Keyboard interactions are commonly recorded as a series of key press and key release events (that is, a sequence in which the user presses a key on the keyboard and then releases it). To create a typing model of the user, different features from those press and release events are generated. In the literature we can find keystroke latency based features, measuring the time interval between two keystrokes [7], [18]–[22]. Keystroke duration is also a common feature seen in research; this feature is measured as the time (from the key press to the key release events) it takes to perform each keystroke [9], [18], [19], [21], [22]. Instead of generating the features from all the keystrokes, some studies aggregate those keystrokes in groups of 2 and 3 keystrokes (called digraphs and trigraphs respectively) and generate the features from these combinations including overlapping key press events [5], [9], [21]. Typing speed was also taken into account as a feature to evaluate [7], [20].

Other studies focus on the use of certain keys when typing to generate predictive features. Some of these keys include error related keys (i.e. backspace and delete) [7], [9], [20], [21], style related keys (e.g. capitalization keys) [7], [18] or other keys such as space bar and enter key [9]. Other features identified in the literature include features related to the position of the keys on the keyboard, such as hand or finger used for typing according to the position of the keys on the keyboard, stylometry (i.e., the measurement of linguistic ''style'', used in authorship attribution and in establishing genre shifts within the work of a single author). Finally some studies ask their participants if they are visual or touch typists, that is if they look at their hands or the screen when typing [18].

Regarding emotion elicitation methods, some studies collect the interactions of their participants while they normally use their computers [5], [7] while other related research proposes guided tasks in order to collect data such as collecting free text [9], [16], using affective stimuli from images extracted from a standardized affective image database, IAPS [23], combined with the input of a fixed text or sequence of numbers [19]. Some works also use data from both approaches [22].

Different approaches are used in labeling emotions from collected data to train supervised learning techniques for affective state prediction. Previous researchers use a categorical approach based on a closed set of proposed emotions. In [7], six basic emotions are proposed to be used by the user to label his or her state after tasks performed in a fixed-text approach and every 15 minutes in a free-texts approach. In [5], users have to provide affective labels using a 5-point Likert questionnaire on 15 affective states for a 10-minute interval of keystroke interactions. In [9] data are labeled with a value chosen from a closed list of emotions after 600 keyboard interactions. Other related studies apply an affective dimensional approach such as [19], which uses the Self-Assessment Manikin scale [12], where participants choose the value that represents the intensity of different affective state dimensions.

It is important to point out that many of these studies deal with relatively small datasets. In [16] Bixler and D'Mello addressed the shortage of data representing certain affective states, modeling only the states that comprised the majority of the affect labels (thus avoiding building models based on few observations). Other studies reported a relatively low number of participants, such as [5] (12 participants) or [9] (9 participants). To illustrate this issue, in Table 2 the dataset sizes found in the research analyzed are shown.

As the research reviewed here indicates, most of the literature on keyboard interaction-based features focus on features described in [5]. However the approach to emotional modeling differs from one study to another in many aspects, including: i) emotion representation, ii) time window (as some studies classify as keyboard interactions at the end of a given task [19] while other studies group interactions on a regular basis given a fixed time interval [5], [7] or a fixed number of interactions recorded [9]), iii) task proposed,

TABLE 1. Features proposed by works analyzed in the related works section.

Data source	Feature type	Application of feature type	References
Keyboard	Keystroke latency	**Keystroke latency (also referred to as "flight time"):** time interval between the key release of the first keystroke and the key press of the following keystroke	[7], [18]–[22], [33]
		N-graph grouped keystroke latency: keystroke latency calculated separately between first and second (digraphs and trigraphs), and second and third (trigraphs) keystrokes of a n-graph	[5], [9], [21]
	Keystroke duration	**Keystroke duration or dwell time:** time interval between a key press event and the release event of that key	[9], [18], [19], [21], [22]
		N-graph grouped keystroke duration: keystroke duration calculated separately for first and second (digraphs and trigraphs), and third (trigraphs) keystrokes of the n-graph	[5], [9], [21]
		N-graph total duration: duration of a n-graph (digraph or trigraph) from 1st key down to last key up	[5], [9], [21]
	Typing speed	**Typing speed:** total number of keystrokes or words per unit of time (minute, second, etc.)	[7], [20]
	Error related features	**Frequency of error:** use of backspace key and delete key	[7], [9], [20], [21], [33]
	Style related features	**Capitalization Rate:** Capital to lowercase character ratio	[7], [18]
	Space, enter, characters, etc. related features	**Frequency of specific keys**, such as enter or spacebar keystrokes	[9], [22]
	Other features	**Sequence**: a list of consecutive keystrokes	[7]
		Hand-based: features according to the hand used to type each key based on "touch-typing" norms	[18]
		Finger-based: features related to the finger used to type each key based on "touch-typing" norms	[18]
		Keyboard row: features regarding key location on keyboard	[18]
		Frequency of common/rare consonant and vowel	[18]
		N-graph number of events: number of key events that were part of the n-graph (digraph or trigraph)	[5], [9], [21]
Mouse	Distance features	**Precision**: relation between distance of two events location (mouse movement start and end, button clicks, etc.) and actual mouse path covered between those two events	[21], [25]
		Distance: total distance covered by mouse cursor	[21], [24], [32]
		Relocation: related features of mouse along the screen (i.e., along x and y axis)	[26]
	Speed	**Speed**: mouse path length divided by time (overall speed, between two button clicks, between pauses, etc.)	[20], [24]–[26]
		Adjusted speed: actual mouse path length between two button clicks divided by shortest path, and then divided by task completion time	[25]
		Instantaneous speed: speed of mouse in different points of the trajectory	[26]

TABLE 1. *(Continued.)* **Features proposed by works analyzed in the related works section.**

Acceleration	**Acceleration: speed change over time**	[26]
	Instantaneous acceleration: speed changes of the mouse in different points of the trajectory.	[26]
Direction/angle	**Direction**: number of mouse movements in a particular direction (In [25] the directions proposed were north, northeast, east, etc.).	[25], [32]
	Angle features: angles described by the mouse trajectory (In [25], where angles are grouped from 0 to 180 degrees by 10-degree step, while in [26] average and sum of angles are calculated).	[25], [26]
Mouse elements interaction	Left/right click frequency: number of clicks	[20], [21], [32], [33]
	Click duration: time between a mouse button press and release	[32]
	Scroll use: number of scroll up or down events	[21]
Pause	**Pause features:** generated from mouse inactivity times (total time, number of occurrences, average time, etc.)	[20], [26], [32]
	Sensitive pause features: generated from mouse inactivity times over a given threshold (0.5 seconds in [26])	[26]

iv) laboratory or real world conditions, or v) stimuli used for emotion elicitation. Other interesting points mentioned in the literature that need to be further explored include verbosity or impact of the topic of the texts presented to participants in the experiments [16], and the differences between fixed and free text approaches [7], [22]. Thus, there is wide variety of criteria when generating features and these do not follow a clear reference framework regarding the analysis of keystrokes. With these in mind, we have designed our study to be carried out in real-world conditions, which avoid practices that are found to have an impact on the task flow, such as writing fixed texts repeatedly or interrupting the task in order to receive emotional stimuli (e.g., through video or audio).

B. MOUSE MOVEMENT ANALYSIS

Mouse interaction has not been widely applied in affective state detection. However, the methodological approach employed with this data source is similar to that was found in our keystroke analysis: i) recording interaction events (traditionally mouse cursor movements and mouse clicks), ii) grouping them and iii) generating features to create a model that varies according to the affective state of the user. Similar modeling issues associated with keyboard data sources appear when modeling mouse interactions such as user skill level (which can be influenced by the device, as the mouse can be a physical device or a track pad) or stress [24].

Regarding the different mouse interaction modeling approaches found in the literature, several features have been identified: i) distance-based features, which are commonly used in most research, ii) precision features (which refer to the relation between the distance in the location of two events and the actual mouse path covered between those two events) [21], [25], iii) distance covered features [21], [24] and iv) relocation-related features of mouse along the screen (i.e., along x and y axis) [26]. Another kind of movement-related feature commonly found in some studies is speed (mouse path length divided by time) [20], [24]–[26], with some variations in adjusted speed (actual mouse path length between two button clicks divided by shortest path, and then divided by proposed time window) [25] or instantaneous speed in different trajectory points [26]. Other movement based features include acceleration and instantaneous acceleration [26]. Regarding the trajectory described by the cursor, absolute direction is used in [25] (proposing directions such as north, northeast, east, etc.) and angle-related features in [25] and [26]. Some studies examine other mouse interactions such as click frequency [20], [21] and scroll frequency [21]. Other related work generates features from the periods of mouse inactivity (e.g., total time, number of occurrences, etc.) [20], [26]. It should be noted that, in order to generate the features proposed, it may be necessary to split the data into small time windows, attempting to adapt these time windows to mouse movements. While keystroke analysis focuses on modeling of keystrokes (i.e., the press and release events of the same key), when it comes to mouse movements there is no clear definition of what a mouse movement is. When trying to identify what a mouse movement is, different studies propose two points of view: some researchers split the data (according to what they consider a mouse movement) every

TABLE 2. Number of participants and data instances from related studies analyzed.

Reference	Participants	Instances per participant	Total instances	Reported top accuracy
[16]	44	3 essays per participant, 1 instance every 15 seconds. Then removed 27.1% instances	1443*	87.0% (Kappa 0.374)
[19]	27	60	1620 (~1460 after cleaning data)	
[18]	1013 (only 486 on both sessions)	10-12 prompts (>300 words) per session		
[18]	704 (2 sets with 352 typists)		4236; 3179; 1416	72.39% (one-tailed Binomial proportion test p P-value: $P<10^{-65}$)
[5]		From 51 to 219 (mean=94.1, s.d.= 52.7).	1129	87.8% (Kappa 0.76)
[9]	9	From 8 to 47	207	81.25%
[25]	44	24	1056*	94.61% (F-Score 0.946, AUC ROC 0.946)
[24]	65	1	65	
[24]	126	1	126	
[24]	80	5	400	
[26]	14			
[27]	367 (133+234)	100?		
[20]	77	10	770 (~600 after cleaning data)	
[28]	55			74.28% (Kappa=0.568)
[32]	96	2	192	
[33]	26	From 21 to 83		
[33]	16	60		
[22]	25			87%

The values highlighted with an asterisk have been calculated from the contents of the paper. Empty cells represent lack of data provided. In those papers where accuracy or other metrics are reported, the values for those metrics are shown.

time the cursor has covered a given distance. For example, [25] splits the data every time the cursor covers 30 pixels. In contrast, other researchers split the data when a pause is found, i.e. there is a period over a given threshold in which no mouse event is registered. The work of [26] is an example of this approach, as [26] splits the data when a pause over 0.5 seconds is found.

Studies using mouse movement for affect detection have had widely varied tasks proposed for mouse data collection. Games where the participant has to identify elements on a screen and click them are common. In [25] participants have to click on 24 different buttons arranged all over the screen. In [27], a visual perception task is proposed, where participants have to evaluate the similarity between different figures. Other studies such as [24] propose using ecommerce websites and product configurators to evaluate the use of the mouse in detecting negative emotions (elicited artificially by means of showing error messages or increasing web page loading times). In educational contexts, some researchers propose the use of an intelligence test designed to be unfair in order induce negative emotion [24], while other researchers use an intelligent tutoring system that contains questions about a previously read text [26].

Both dimensional [24] and categorical [25] approaches are used to label emotions, as seen in keyboard studies. The difficulties of using any given labeling approach are illustrated in [25]: the researchers initially proposed Russell's circumplex model (a dimensional approach for emotional labeling) but rejected this model in favor of a 7-point Likert scale (where 1 = content, and 7 = confused). Preliminary results revealed that participants were only able to label two categories of emotions (content and confused).

The number of participants in mouse-related works, range from 14 participants [26] to more than 300 [27] (detailed numbers can be found in Table 2). Differences in experimentation settings give rise to a disparity in the number of participants involved, which along with differences in the number of instances per participant (see Table 2) have an impact on modeling issues and obtained results, as discussed below.

As a result of this review, it can be seen that most research using mouse movement analysis in affective states detection rely on different sets of features from interaction data. There is a lack of standardization in some basic concepts, such as ''mouse movement'', which is defined in different ways (straight lines of 30 pixel length in [25], or a concatenation of coordinate changes in the cursor with no more than 0.5 seconds between them [26]). Some studies pointed out the impact of using context information in predictive models [25] or considering prevalence of an emotion and the persistence of that emotion over time after the presentation of a given stimuli [26]. Due to the differences found in the review of research related to mouse analysis, it can also be concluded here that there are no consistent, general approaches nor a clear reference framework of features. Lastly, the implications of this present study may likely benefit real-world students.

C. MOUSE MOVEMENT AND KEYSTROKE ANALYSIS

Past work has combined both data sources in affective state detection. These have followed approaches that may take into

account not only issues related to each data source separately, but also features that may arise from the combination of both, such as the ratio of interaction with each data source, pauses when switching from one source to another, etc.

Regarding the features used, there is a combination of features already presented in previous subsections (as discussed in [28]). The inclusion of other data sources is also proposed such as voice [29]. In [30], for instance, voice features are combined with keyboard and mouse features.

Upon closer examination, researchers investigate mouse movement and keystroke analysis in the context of the educational field and e-learning platforms [30] with diverse tasks such as mental arithmetic tasks [20] and programming courses [28]. Specifically, researchers propose clustering happy and neutral emotions that encourage learning, and anger, sadness and fear as negative emotions that discourage learning [30]. Some studies focus only on specific negative states (i.e., boredom, confusion and frustration) as these have an impact on student achievement and learning [28], [31]. Other studies consider learning-related goals such as cognitive stress [20]. Another common approach in this field is to use task performance-related features, ranging from general task-related features [20] (e.g., task completion time, error of tasks, etc.) to more specific features that depend on the nature of the task being performed (e.g., the use of keys to run a program in programming courses [28]).

With regard to the usefulness of combining mouse and keyboard, [28] generated different models, with keyboard features only and with a combination of mouse and keyboard. Optimal results are obtained when both data sources are combined. In [32] Zimmermann *et al.* also propose the use of mouse and keyboard features to detect affective states in an e-commerce shop using film clips, toward the overall goal of inducing different states. In [33], two studies based on mouse and keyboard use are introduced, using emotional elicitors such as music and physiological signals (galvanic skin response). These findings indicate that it is possible to measure individual moods based on keyboard and mouse interaction, but these results do not hold true when creating a generic mood predictor based on the same data. Both [32] and [33] use a two dimensional approach based on self-reported valence and arousal ([32] discretizes the dimensional values into categories while [28] using a categorical approach, generating the annotations from video segments).

Prior work was also presented in previous state of the art analysis [6], which provides a comparison of which data sources were used, emotion elicitation methods, emotions analyzed, data labeling methods and application fields among other variables. Another state of the art analysis can be found in [34], with an emphasis on the educational field. Like our study, this paper discusses the difficulty of establishing clear points of comparison within the literature, highlighting the lack of common methodological elements.

As we can see, there is scarce research combining data gathered both from keyboard and mouse. Additionally, most of the main points discussed in previous research are tackled from different angles (such as the approaches used for labeling emotions, introducing clustering emotions [30], or different emotion elicitation methods such as music [33]). Overall, studies have also shown that the combination of both data sources provide better results [28].

D. FEATURES OVERVIEW

From all the works reported in the previous subsections, Table 1 includes all the features (or groups of features) that have been identified. These features are grouped per data source and described according to the varied ways these have been applied in research.

In a previous work on affect state detection through keyboard and mouse interaction devices (see [21] for details) we already used several features which are included in Table 1. These features, along with the new ones that are used in the study are introduced in section III.B.1.

E. KEY MODELING ISSUES

The outcomes of the review of studies using keyboard and mouse interactions points to a number of modeling issues that have not been properly addressed. Notice that we paid special attention to modeling issues that were already introduced in Section I (e.g., intrusiveness, emotional characterization, context of the experiment for data collection, etc.). The methodological differences that can be found in literature to address those issues are as follows:

1. **Emotional elicitation method used:** in order to obtain a large number of instances to train the models, many approaches are based on eliciting emotions from their participants. The emotion elicitation methods that were found include watching videos [19], faking slow load times or errors [24], dealing with adjustments in an intelligent tutoring system [26] or performing everyday computer tasks [5].
2. **Task proposed for data collection:** as to keyboard interactions, some works proposed a fixed text approach [19] while others get data from a free text approach [9]. Although some papers have explored both approaches [5], [22], free text results seem to provide worse results (in [5] there were no free text models achieving an overall classification rate $> 75\%$, while in [22] free text models provide lower accuracy rates than fixed text).
3. **Temporal applicability of the model:** some papers use a temporal threshold in order to define time windows to model the user interaction (e.g. 15 minute time windows in [16] or 10 minute time windows in [5]), other work propose a number of interactions threshold (e.g. 600 keyboard interactions in [9] or 30 pixels mouse movements in [25]). Notice here that most studies do not justify the thresholds used. This is an important point as the modeling performed to the interaction performed is the main source of information here presented. Also should be taken into account how the proposed task impacts on the temporal applicability of the model, as, in the cases where the

participant might not use some kind of interaction device for a long time, it would be interesting using the number of interactions threshold.

4. **Interaction modeling features:** the skill of the participant when dealing with the proposed interaction device has rarely been taken into account. While some related features were considered in previous work (i.e. hand-based, finger-based or looking behavior when typing in [18]), most studies ignore the impact of the users' skill on the models. This is an important point that may impact interaction behavior patterns, especially in between-subject approaches where data from different users will be used to train models. Ignoring this issue, might force to model the participants by their skill but not by their own interaction changes over time, which aims to be studied in this work.
5. **Affective labeling approach:** Labeling affective states to train data mining models has been a long-term issue in affective computing research and, especially, in education modeling [35]. There has been extensive work on using experts' knowledge to label users' affective states [9], [36], with increasing success even in natural settings (e.g., students in a school) [37]. Some studies use a categorical approach, with different categories depending on the approach (e.g. 6 basic emotions were proposed in [7] whereas only two emotions were evaluated in [25]). Other work, in contrast, follows a dimensional approach [19]. Also the labelers of affective data is a point of discussion, as self-labeling can be the only alternative where trained expert labeling is not available (e.g., at home).
6. **Number of data instances used and class distribution:** gathering data with this affective prediction purposes requires the design of experiments where many users are involved. Many researchers generate models using datasets which are based on many features but with few instances available, as in the case of modeling from real world experiences,. This is commonly known as the curse of dimensionality, and should be avoided to keep from generating overfitted models. Another point that results from the experimental design followed is the class distribution of the dataset. One of the goals of a good experimental design is to collect sufficient instances of all the classes, in order to avoid unbalanced datasets. There are experimental design issues that take part in the class balancing of the dataset, such as the elicititation method used, which should be tailored to the experimental subjects. To mitigate this problem data preprocessing techniques can be applied. Although the number of data instances used in the model generation is not clearly described in most works, Table 2 illustrates (when possible) the approximate size of the datasets drawn from the different sources found in literature.

Other issues should also be taken into account. In particular, as introduced in section I, our approach is framed in a real-world educational scenarios and therefore some basic performance features are considered. Thus, we have focused our attention on how previous studies deal with similar performance features. In this regard, [16] and [18] along with [26] are the only studies that propose education-related tasks and involve keyboard and/or mouse with affective purposes. Across these studies, there is no evaluation on how signals may impact on learner outcomes, only affective state detection. Nevertheless, there is strong evidence on the connection between affective states, learning [38] and motivation [39].

To conclude, in this research we focus on the last three issues on the list, considering the skill of the participant, affective labeling and the application of data preprocessing techniques (in order to deal with small and imbalanced datasets) as experimental variables.

III. PROPOSED APPROACH

Although we have conducted previous research on detecting affective states from keyboard and mouse interactions [21], this was based on a much less sophisticated approach than the one we present in this paper. Here the focus is on evaluating the impact of specific methodological issues found in the affective computing field as described in section II.E. In the present paper, we proposed to study methodological questions, including intrusiveness (both of the data sources used and of the task employed to elicit or label the affective states), emotional modeling, data preprocessing techniques used and the inclusion of an interaction baseline model. We have developed our methodology as indicated in the literature reviewed in Section II: i) collect data, ii) provide affective labels for the data, iii) prepare the data, and iv) generate predictive models. Our lab work on these issues [40]–[43] did not take into account the methodological issues that we are considering here regarding data preparation, but it has provided us the support and methodological experience required to conduct experimentation that is described in this paper.

In order to generate a baseline model, a reference text that is task-independent is usually used to get an individual interaction pattern which compares a user's performance over different tasks [5], [7]. The choice of text may have an impact on the quality of the initial model in different ways and we need to consider two variables: text length may impact the usability of the approach (long texts are to be avoided as they may distract learners from their regular learning task), the number of different words used in the text may alter the richness and variety of features in the model [16]. Another issue is which time window to use in detecting emotional state changes. Here, the approach depends on the given task and can be either a time window (e.g., a 10-minute window of keystroke interactions is used in [5] and [33]) or a number of events recorded (e.g., 600 events in [9]). We are also taking into account performance features related to the learning task because these may have an impact on the participant's affective state as reported in [20] and [25].

The current approach extends our prior work in four ways (see Fig. 1). First, we have defined a baseline model (with a resultant user-normalized dataset represented in bottom left part of Fig. 1), which refers to how the participant generally interacts with the keyboard and mouse (as the typing and

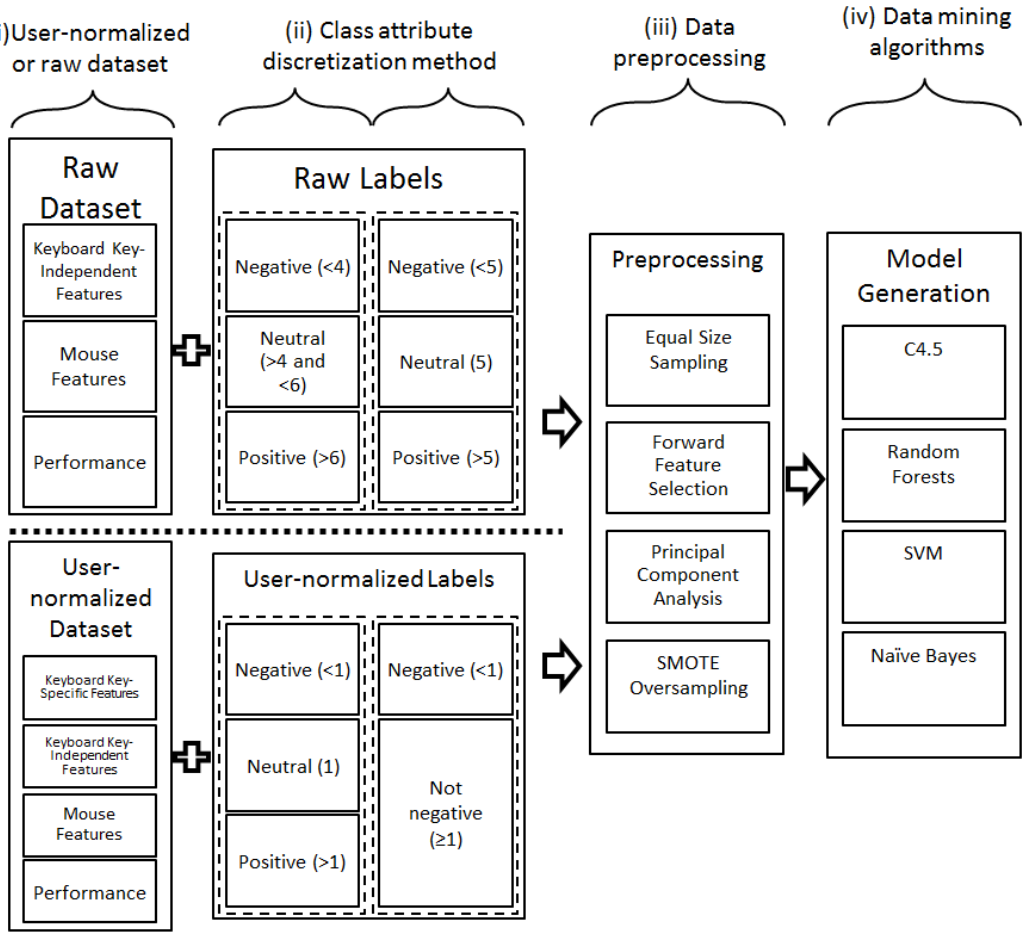

FIGURE 1. Methodological aspects evaluated in this paper: i) comparing user-normalized dataset and raw dataset, ii) comparing different approaches to discretize the affective labeling, iii) different preprocessing techniques used with the data, and iv) the data mining algorithms to be used.

mouse behavior could also be affected by the user skills). This prevents us from dealing with different typing skill level for each user, the fourth open issue pointed out in section II.E. To create this model, an initial calibration task has been included to collect reference values used in generating keyboard and mouse interaction features, as well as the affective labels. This is a task in which the user is not affectively challenged and the goal is to obtain an initial model of each user's interaction behavior. In our experiment we evaluate the use of this calibration task. We record the way the participants interact during an initial task (the calibration task) and then, we evaluate what interaction changes occur during the different tasks over the experiment. This provides us with a user-normalized dataset (as the data is normalized using the initial data recorded for each user). Another data set is obtained from considering only the participants' actual usage over the course of the experiment, i.e., excluding the calibration task. This is called the raw dataset. Second, we process the emotional labels, using different discretization approaches (in both dimensions proposed: valence and arousal) while addressing the third issue mentioned in section II.E. Third, we apply commonly used dimensionality reduction and class balancing techniques to process the data from high dimensionality or unbalanced data sets (sixth open issue in section II.E). Fourth and last, we use multiple data mining algorithms to generate the model.

Our previous research [21] did not require the use of the keyboard until the end of the experimentation tasks, when the participants were asked to write about their emotions while they were involved in solving problems. In this new study, two main changes were introduced in that direction: i) three keyboard-centered tasks were proposed in order to collect more keyboard interactions (although the mouse can be used for text editing and is mandatory to navigate between the tasks) and ii) the subject of the text was changed to replicate an educational task, which consisted of writing an essay as part of an English as a Second Language class. These changes provide more keyboard interactions, thus enabling the creation of a more robust user model from a larger dataset in an educational context.

Our approach takes into account key modeling issues: preprocessing techniques for class balancing and dimension-

ality reduction, discretization method applied to the affective label attribute, feature individual normalization across users' interaction behaviors and data mining algorithms used. To cover these issues, this section describes in detailed both the process followed to collect the data (subsection III.A) and the different feature generation processes followed according to the different proposed data sources (subsection III.B). The process followed to collect the data includes the experimentation setup with materials and participants. In turn, the feature generation process includes the datasets for both mouse and keyboard interaction features along with the task performance and predicted attribute.

A. EXPERIMENTAL SETUP AND DATA COLLECTION

An experiment was designed taking into account the following points that have been discussed in previous sections: i) addressing a real-world educational scenario, ii) taking into account low intrusiveness in terms of interaction devices and experimentation settings, iii) accounting for individual's behavior patterns, and iv) following an affective self-labeling approach.

As our goal is to evaluate this approach in a real-world scenario, a real-world task was chosen. In order to collect interactions with using a keyboard, an essay writing task was designed (similar to those in previous work [16], [18]), where keyboard interactions are mandatory and mouse interaction would be needed to edit the text as well as to navigate through the application. An emotional elicitor is also needed to record different affective states from participants. Standardized emotional stimuli employed in other studies such as sounds [44] or images [23] were not used in our research as they are not present in real-world educational scenarios. Instead, task difficulty was chosen to elicit different emotions from the participants [45], [46]. English as second language in a classroom was chosen as the context for our experiment, as it enables us to manipulate the desired difficulty level of the materials within the context of the scenario itself. Our study attempts to take advantage of both within-subject and between-subject approaches to data collection and analysis. On the data collection side, our naturalistic, between-subject experiment would generate few data instances from each participant, which makes it more difficult to get an accurate set of features from each participant than in within-subject experimental approaches, in which more interaction data are considered [9]. With regard to data analysis, an initial calibration task is proposed in order to generate an initial model for each participant, to be used to evaluate individual keystroke, mouse movements and affective mood changes over the tasks (following a between-subject approach).

An essay writing tool called MOKEETO (MOuse and KEyboard logging Essay writing TOol) was implemented in order to log all user interactions with their corresponding timestamp. The tool consists of a sequence of panels (each one corresponding to a different task) with three main sections: i) the task instructions shown on the top of the screen, ii) a text input form in the center of the screen and iii) a set of indicators (the number of written words by the participant and a timer showing the amount of time left) at the bottom of the screen (see Fig. 2). Additionally, each task can be configured to: i) show in the task instructions section (without allowing copy/paste) a set of words to be included in the essay, ii) hide

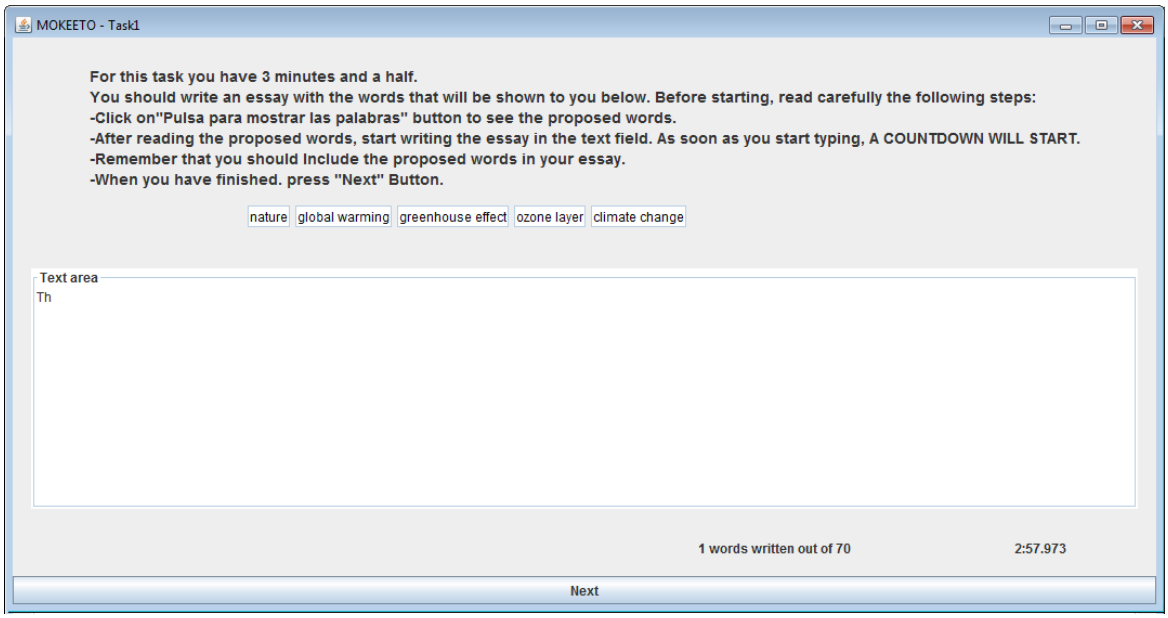

FIGURE 2. MOKEETO essay tool used in the experiment. Instructions are given at the top of the screen, proposed words are shown below task instructions and the text area is shown in the middle of the screen. Word counter and time remaining are shown at the bottom of the screen.

that set of words after a given time and disable the text input form while the words are being shown (here the participant has to remember the words before starting to type the essay, which may increase the difficulty of the task), and iii) prevent the participant from ending the current task until a set word count or time limit had passed (which remained across the set of tasks).

The set of tasks proposed were as follows. An initial task (referred to as the ''calibration task''), which involves asking the participants to copy a short text from Alice in Wonderland (as seen in [5] and [7]), was designed to create a baseline model of how the participant interacts with the keyboard and mouse. This baseline model can be used as a reference when comparing the keyboard and mouse interactions of the next tasks. In previous research, this baseline model has also been traditionally used in experiments using physiological data sources, and our approach has adapted this idea to our current sample of interaction data sources. The usefulness of this baseline model is the key methodological question we aim to evaluate (see hypothesis H in the introduction section).

Because the purpose of our research is to detect participants' affective changes, and bearing in mind the relationship between affect and cognition demands discussed in previous related research, several factors were used to increase the difficulty of the tasks over time. These include: i) time limit with a countdown (as used in [20]), ii) incremental difficulty of proposed words (proposing uncommon words in the last task and common words in the first tasks), and iii) forcing the participant to remember the required proposed words that need to be used in the written composition (as in the last tasks the proposed words are only shown for a limited time before the participant is allowed to write). With these factors in mind, the following regular tasks were proposed (in addition to the initial calibration task). The first task consisted of writing a text with 5 proposed words that were shown throughout the task. The second task was like the first one, but the proposed words were shown only 30 seconds before the participant was allowed to start writing the essay, forcing the participant to remember the 5 new proposed words during the task. The third task was similar to the second one, but the words presented were highly uncommon. All the tasks had a time limit of 3.5 minutes (210 seconds), with a visible countdown.

After each task (including the initial calibration one), participants were asked to express their affective state at the end of each task. Self-report is one of the most common approaches to determine affective states as seen in most research analyzed in [6]). The self-report selected for our experiment was the Self-Assessment Manikin (SAM) scale (a standardized way to label emotions in a dimensional way [12]), providing a score, from 1 to 9, for both valence (indicating the extent to which the affective state is positive or negative) and arousal (indicating the strength or activation of the affective state) dimensions. This way of reporting affect was chosen to avoid working with a set of different predefined emotions and focus on changes on affective dimensional values. The SAM scale was presented with a textual explanation of each one of the affective dimensions to label. Valence and arousal values will be used to generate the affective attributes to be predicted by the system.

1) MATERIALS

The hardware configuration consisted of a laptop connected to a USB keyboard (Dell Sk-8115 with Spanish keyboard layout), a USB mouse (two buttons with a scroll, Logitech M100) and an external monitor (TFT 15'' Dell E152FPB).

Together with the MOKEETO essay writing tool, a key logger and mouse tracker developed in this research (and previously used in [21]) were running in the background, tracking of all the press and release events registered by the keyboard as well as the movements and clicks registered by the mouse. The vocabulary for the tasks proposed was extracted from the materials being used in English as second language classes for a B2 CEFR level.

2) PARTICIPANTS

Two similar experiments were conducted: the first one in April (with 27 participants, 10 males and 17 females, avg. age 15.41) was held in a high school class. The second one in November, with the same setup, was held in a different context, with participants attending an experience organized by our distance learning university (UNED) within the Madrid Science Week (14 participants, 7 males and 7 females, avg. age 44.35). The materials in both experiments were the same and aimed to replicate a real-world educational scenario of foreign language learning.

B. DATA PREPROCESSING

After the experiments were conducted, all data collected had to be processed in order to create the model of the participant's interactions. These participant interaction features were inputs for the data mining algorithms used to detect the learner's affective state. The time window corresponds to the duration of each task, so for every user and task a vector was generated, which contained features representing the interaction of the user during that task. By using this time window, each generated vector has a corresponding target or label attribute to be predicted (i.e., the participants' self-report of their affective state after each task).

1) INTERACTION FEATURES

This subsection describes the different datasets generated from data collected during the experiment to model participant's interactions. To model keyboard interactions, key events have been recorded (key press and key release time) and processed, generating two different sets of features, with key-specific features and key-independent features. To model mouse interactions (specifically mouse movements), click and scroll events were recorded. Note that although the latter was not used as the graphical user interface, the use of scroll was not actually required, generating only one set of features. To model performance, a particular single

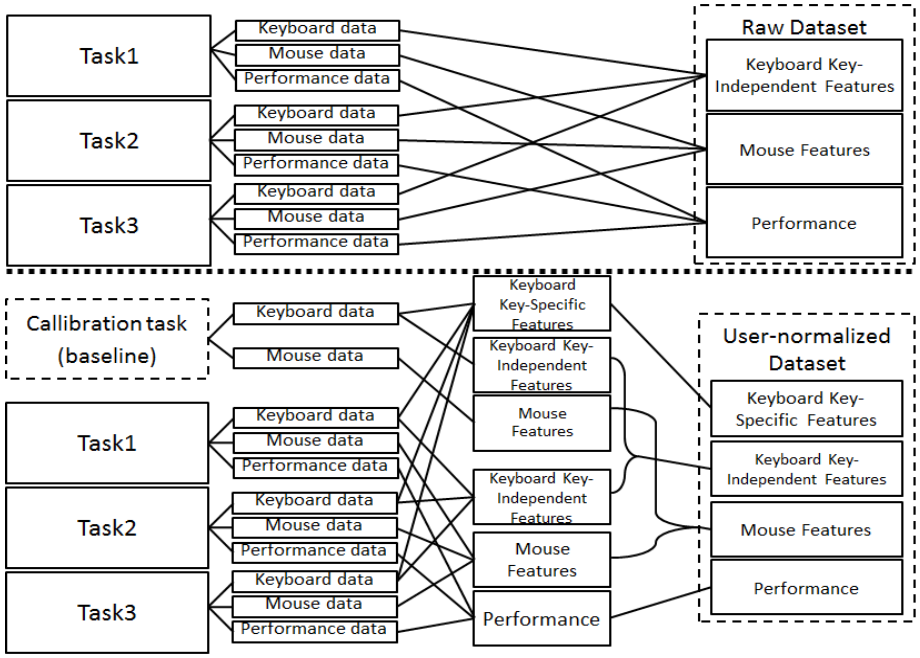

FIGURE 3. User-normalized and the raw datasets generation process differences. User-normalized Dataset (bottom right) is generated by comparing data collected from the calibration task and the regular tasks. Raw Dataset (top right) is generated using only data collected from the regular tasks.

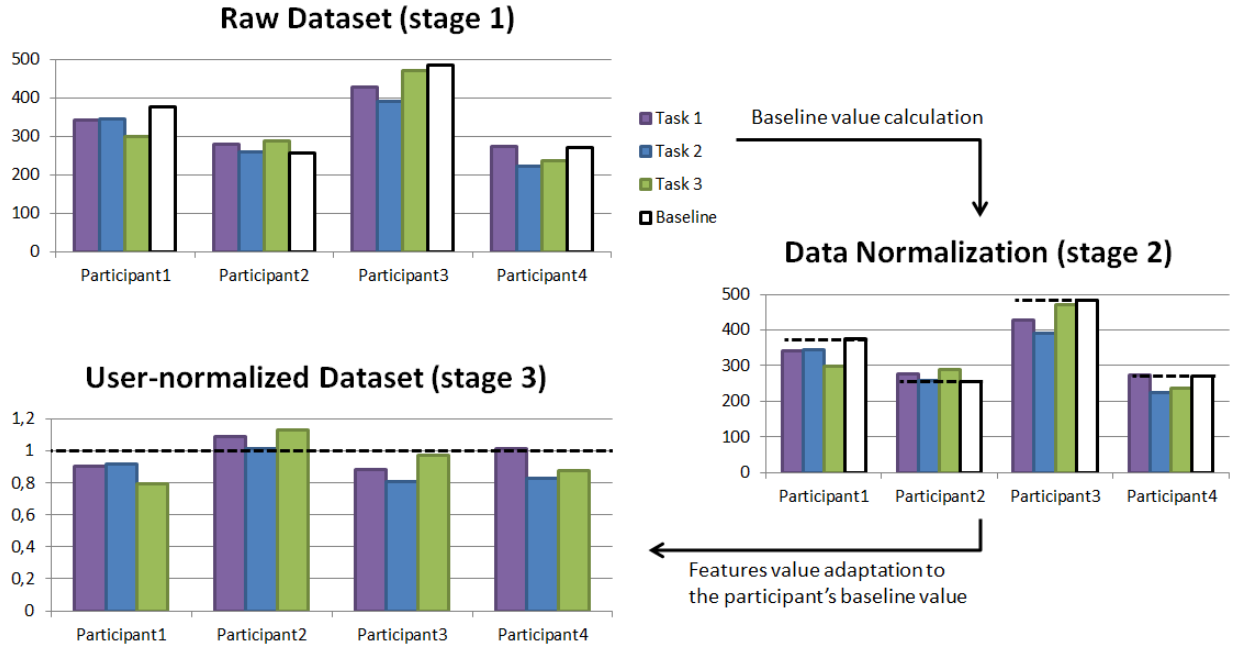

FIGURE 4. A graphical description on how the user normalized values are calculated (showing 4 out of the 41 participants). Slashed line represents the reference value measured in the baseline task for data normalization for each participant in stage 2 and the reference value after the normalization in stage 3.

set of features was generated. To model the user affective state, different approaches to transform the participants self-reported labels into different affective categories (described below) were used and compared. For the mouse, keyboard and affective labeling datasets, two versions were created: one using information from the calibration task baseline (i.e., user-normalized dataset) and another without using that information (i.e., raw dataset), as it is illustrated in Fig. 3. The purpose here is to evaluate to what extent using a baseline to normalize regular tasks improves data prediction models.

The process followed to generate the user-normalized dataset feature values is as follows (values and process are shown in Fig. 4). Initially, the raw data (Raw Dataset in stage 1) is obtained, where each participant has a particular skill

level using the keyboard and mouse that is represented by a specific value. In the Data Normalization stage (stage 2), the values collected in the calibration task are used as a reference value for each user. Each value recorded for a participant in every task is divided by the value of that specific participant in the baseline model (i.e., in the calibration task). The result of this is used in the final step to build the User-normalized Dataset (stage 3), where the goal is to get rid of the possible differences in the values among participants due to their respective keyboard/mouse interaction skills. In other words, in stage 3, values represent the proportion a feature has changed compared to the baseline value (for that same feature for that given participant). The normalized values were calculated using the following equation:

$$Normalized\ value = \left(\frac{Raw\ value\ calculated\ in\ task_x}{Raw\ value\ calculated\ in\ calibration\ task}\right)$$

a: KEYBOARD KEY-SPECIFIC FEATURES

This dataset aims to model the differences reflected when typing the same sequence of keys in a task compared to the way the same sequence of keys was typed in the initial calibration task. To generate this model, the first step was to group all the combinations of two and three consecutive keystroke events in every task (including the calibration task). Once all the combinations were created and in order to generate a precise model of the user's typing, only combinations that were typed during the same task over a given number of times were kept (i.e. the sequences of the same keys that were introduced over a given number of times in the same task). In contrast with [18], where no minimum number of instances per n-graph is set, we take into account n-graphs that have been typed at least 3 times across all tasks, including the calibration task. This decision was made in order to mitigate the impact of possible noisy n-graph instances, toward the goal of generating only robust models from several observations of the same phenomena. Fig. 5 illustrates the impact of this threshold in the numbers of digraphs generated for each task and the number of digraphs used in common with the calibration task. The more instances each n-graph are required for it to be used by our model, the less number of n-graphs meet that requirement. Fig. 5 illustrates the average number of digraphs present in each text and the average number of digraphs used (i.e., the number of digraphs that appear over n times in the same text).

For these combinations, all the digraph and trigraph features present in [5] were calculated for each task (includ the calibration one) following the same coding presented there:

- 2G_1D2D: The duration between 1st and 2nd keypress start times of the digraphs.
- 2G_1Dur: The duration of the 1st keypress of the digraphs.
- 2G_1KeyLat: Duration between 1st keypress end time and next keypress start time of the digraphs.
- 2G_2Dur: The duration of the 2nd keypress of the digraphs.

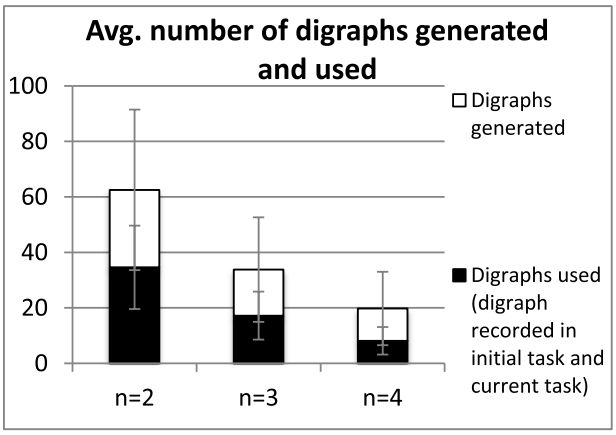

FIGURE 5. Avg. number of digraphs generated for each task depending on the number of instances (n) of that digraph in the text. Also, number of digraphs (of the same type as those recorded in the calibration task) used for each task.

- 2G_Dur: The duration of the digraphs from 1st keypress start time to last keypress end time.
- 2G_NumEvents: The number of keypress starts that were part of the graph (as possible overlapping may occur).
- 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_2D3D: 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.

Once we have all the features for each task, we calculate the proportion for each feature compared to the same feature and same digraph in the calibration task, following the equation presented in Section III.B.1, using that proportion for each feature in the dataset.

b: KEYBOARD KEY-INDEPENDENT FEATURES

The second dataset generated was based on key-independent n-graphs key-related features (i.e. all the n-graphs are aggregated, regardless of the keys that compose them). As described above, the features presented in [5] were also generated in this dataset. The main difference with the keyboard key-specific feature model is that, in this dataset, all the n-graph features are aggregated. In the keyboard key-specific feature model, only a subset of the n-graphs recorded was used (only those that appeared over a given threshold in the given task and the calibration task).

Also some other features not included in [5] were generated in this dataset:

- Overlapping press events: number of press events occurring while another key was already pressed.
- Uppercase press: number of press events occurring while shift key was already pressed.
- Pauses: time between a key release and a key press events.
- Time between two consecutive press events: time between two different consecutive key press events.
- Time between press and release events: time between the key press and key release events of the same key. This indicator was calculated taking into account different sets of keys (backspace key, backspace and delete keys, delete key, alphabetical characters and space bar).
- Word separation: time between the release event of a character key and the press event of another character key separated with a keystroke of the space bar.

For this dataset, two different versions were generated: i) the ''raw'' version of the dataset containing values for each feature calculated in each task, and ii) the ''user-normalized'' version of the dataset containing, for each user, the comparison between the feature values in each task and the corresponding feature value in the calibration task following the equation presented in Section III.B.1.

c: MOUSE FEATURES

The third dataset was generated from the mouse interaction logs (containing mouse movements, clicks and scroll movements). The first step is using an aggregation method to group the events from the raw data recorded to generate the model features. Traditionally, in mouse interaction modeling, events are grouped into mouse movements; however, as previously discussed, there is no standard definition of mouse movement. In [25] the mouse events were grouped into segments or movements based on their distance, in [26] is the time of inactivity that is used to separate mouse movements. We built our model following the second approach and all the mouse cursor coordinates recorded during the experiment were grouped into mouse movements, a series of coordinates that vary along time with a time difference between each one below a given time threshold. If a change in the cursor position is produced after a 500 millisecond inactivity time (as in order to have a base already proven in related woks, the threshold was set to 500 milliseconds as it was done in [26]), that position will be considered as the starting position of a new mouse movement. After that, the following features (proposed in [21] and [27]) were calculated:

- Movement accumulated angle variation: the angle variation described by the cursor for every pair of consecutive cursor locations compared to the angle described by the previous pair of cursor locations.
- Average movement acceleration: the average acceleration in each movement.
- Movement acceleration standard deviation: the standard deviation of the acceleration in each movement.
- Average movement speed: the average speed of the mouse movement.
- Movement speed standard deviation: the standard deviation of the mouse speed in a mouse movement.
- Distance covered: the distance covered by the cursor in a mouse movement.
- Euclidean distance: Euclidean distance between the coordinates where a mouse movement begins and the coordinates where the mouse movement ends.
- Difference between ''distance covered'' by the mouse cursor in a movement and ''Euclidean distance'' between the starting and the end point of the movement.
- Click covered distance: the distance covered by the cursor while a mouse button was clicked.
- Click Euclidean distance: Euclidean distance between the coordinates where a mouse button was pressed and the coordinates where the mouse button was released.
- Click time: time between a mouse button press event and the consecutive mouse button release event.
- Difference between ''click covered distance'' and ''click Euclidean distance''.
- Pause length: the length of the pauses between mouse movements (>1s).

Similar to the key-independent dataset, two different versions, user-normalized and raw, were generated from this dataset in order to evaluate if the use of the calibration task data for each user impacts predictive outcomes.

2) TASK PERFORMANCE FEATURES

A dataset reflecting the performance of the participants during the experiment was also generated. This way, since we are modeling affect in the educational domain, we can utilize a performance model to improve the prediction of whether the users' performance has an impact on their affective state. This approach was discussed in the previous work [20], [25], where authors used the best trajectory in a task to evaluate the participant's performance. Bearing in mind that proposed tasks consisted of writing a short essay including certain selected words, the following features were included:

- Proportion of words used in the task compared with the mean number of words written by all the participants in that task.
- Proportion of proposed words used in the task.
- Proportion of proposed words in the task used compared to the mean proportion of proposed words for that user in other tasks.

Previous work also use verbosity [16] or other language production features [18], which may be closely related to the domain of the language used. In contrast to these studies where participants were American students, our work required students to write essays in a second language. Domain performance features relating to language production such as verbosity were not considered in the present study, as these features are likely not only associated with a participant's current affective state but also associated with their proficiency in the second language.

3) LABELLING TARGET AFFECTIVE FEATURES (PREDICTED ATTRIBUTE)

The last dataset generated consisted of the label to be predicted. As aforementioned, the approach chosen for this work is based on a dimensional approach so the labeling does not consist of concrete individual affective states. The affective labeling used is based on the two affective dimensions the participants were asked after each task (with values from 1 to 9 both of them): valence and arousal.

Related work using the same labeling scale group the values given by participants into categories [47], merging both valence and arousal dimensions into one single variable (with values: PVHA, PVLA, NVHA, NVLA and nVnA where P=positive, N=negative, H=high, L=low, n=neutral, V=valence, A=arousal). Unfortunately, [47] did not give detailed information on the procedure used to carry out that discretization.

In our case, we consider the discretization process of the participant's self-reported labels, searching for a balance between fine-grained modeling (the label given by participants) and prediction performance and simplicity. As with fine-grained modeling, the worse the results might be (as a more detailed prediction is expected) and more complex the model). For this dataset, two versions (user-normalized and raw) were also generated aiming to design two different targets to evaluate: i) values given by participants after each task (raw version of the dataset), and ii) comparison between the values given by participants after each task and the values given at the beginning of the experiment (user-normalized version of the dataset).

To model those positive or negative absolute numerical values (i.e., labels) given by the participants (raw version of the dataset), two different discretization approaches were applied: i) the first discretization approach aims to draw a narrow neutral strip in the affective dimensions, considering positive (>5), neutral (5) and negative (<5) categories, while ii) the second discretization approach aims to predict strongly positive (>6) and negative (<4) scores, with a wider neutral range (>4 and <6).

Recall that participants reported their affective state before doing the first essay task. Thus, using the user-normalized version of the dataset allows us to know if the participant has gone through a positive or a negative emotional transition since the beginning of the experiment (as the user-normalized version of the dataset uses the initial emotional labeling given by the participant as reference values). As done in the raw version of the dataset, two discretization approaches were also applied (this time with the values given by participants compared to their own baselines values obtained in the calibration task): i) a model including a neutral value, which includes positive (>1), neutral (1) and negative (<1) categories, and ii) a model focused on detecting negative transitions, which include only negative (<1) and non-negative (≥ 1) categories. These two different discretization approaches are depicted and highlighted in dash line boxes in Fig. 6. The latter

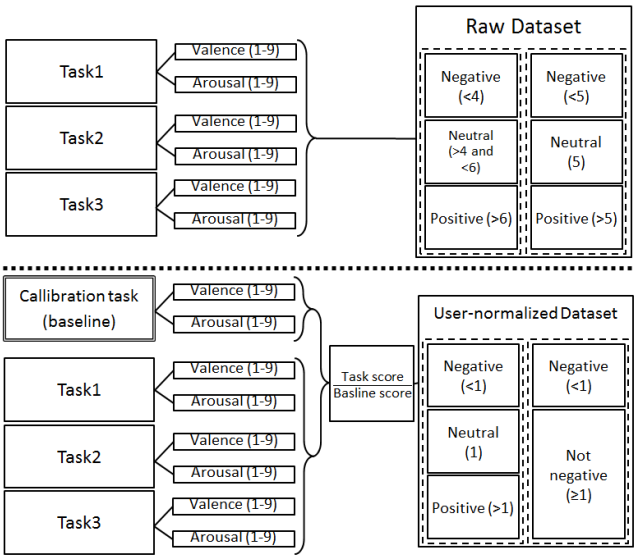

FIGURE 6. Different discretization approaches used for the labeling target affective attribute. Each different approach is included in a dash line box. Differential labels (bottom part) are generated by dividing labels given in regular tasks by labels given in the calibration task. Raw labels (top part) are generated from the labels given in regular tasks.

negative and non-negative approaches are being explored here because these have been considered in our previous research on recommender systems [48], where the recommender system may only provide a recommendation when the learner is going through a negative affective state.

C. DATA ANALYSIS AND PREDICTIVE AFFECTIVE MODEL GENERATION

After discussing the different ways to consider the features and labels in our datasets (keyboard key-specific dataset, keyboard key-independent dataset, mouse interaction dataset and performance dataset) in this section we focus on describing how we generate the predictive models that are ultimately tested.

As it can be seen in Fig. 3, two versions were generated from most of the datasets previously discussed (i.e., user-normalized dataset, bottom part in Fig. 3 and raw data set, top part in Fig. 3). Both versions of the datasets contain a combination of the feature vectors from each dataset, linking them to the same set of events, and merging the tables. The only exceptions were the keyboard key-specific and the performance datasets. This is because we process keyboard key-specific features by comparing its values to the calibration task. In the case of the performance dataset, there are no reference values from the calibration task. Both versions of the datasets, raw and user-normalized, were used separately to generate different models. By doing this, we investigate how "user-normalized" feature values (using the baseline data from the calibration task to generate the features) or "raw" feature values (not using the baseline to generate the features) impact the quality of the predicted models.

When all the datasets are merged, we found that we had a higher number of features than instances: 406 features for 123 data instances for the ''user-normalized'' version and 258 features for 123 data instances for the ''raw'' version. To address this dimensionality issue, a reduction was needed. After removing columns and rows with a large number of missing values, highly correlated features were also removed. To do that, features were ranked by number of correlated features and then those features that correlate with another feature with a higher rank are removed. After this step, high dimensionality was still present: 165 features for 119 data instances for the ''user-normalized'' version and 95 features for 119 data instances for the ''raw'' version. As a consequence, dimensionality reduction and class balancing were needed and their impact on the results was also evaluated (these are to address the sixth issue depicted in section II.E).

The techniques that were applied for class balancing consist of: i) equal size sampling, which aims to balance the class values in a dataset with an uneven class distribution (problem that can be found in previous work [9]) as that may lead to overfitting ([49]; and ii) SMOTE, which performs oversampling on the minority class examples [50]. In turn, the techniques that were used for dimensionality reduction consists of: i) forward feature selection, which selects a subset from the original set of features that provide the best predicting results [51], and ii) Principal Component Analysis (PCA), which, after having removed correlated features, provides an uncorrelated set of features generated from the remaining ones, allowing us to select a subset of features with the most plausible discriminant dimensions [52]. For the last methodological variable, we consider which algorithm will be used in our study based on the results obtained with our given data. As such, we have used some of the most popular applied algorithms in the literature [6]: i) C4.5, an algorithm to generate decision trees [53], ii) random forest, an ensemble learning method based on the creation of multiple decision trees [54], iii) support vector machine (with a polynomial kernel), another supervised learning technique based on the construction of hyperplanes for splitting the data points in order to assign them a class value [55], and iv) Naïve Bayes classifier, which is based on the probability to perform the classification [56]. The models were generated using the Weka implementation in the Knime platform [57].

The final list of the different methodological variables evaluated in both datasets used in this affective model prediction is shown in Table 3.

TABLE 3. Methodological variables evaluated in this work.

Variable	Values evaluated	
Label or class attribute	Valence Arousal	
Label or class discretization method	User-normalized dataset	Positive(>1), Neutral(1), negative(<1) non-negative(≥1), negative(<1)
	Raw dataset	Positive(>5), neutral(5), negative(<5) Positive(>6), neutral(>4 & <6), negative(<4)
Equal size sampling	Yes No	
Forward feature selection	Yes No	
Principal Component Analysis	Yes No	
SMOTE oversampling	Yes No	
Algorithm	C4.5 Tree (J48) Random Forests (RF) Support Vector Machines Classifier (SMO) Naïve Bayes (NB)	

IV. RESULTS

With all the variables described in the previous section, 512 models were generated, all of them applying 10-fold cross validation. The results of the models with best accuracy, best kappa and best accuracy improvement (compared to the majority class) are shown in Table 4 (user-normalized dataset) and Table 5 (raw dataset).

Top results are bolded. Settings indicate which preprocessing technique is used: equal size sampling (D), forward feature selection (F), principal component analysis (P) and SMOTE oversampling (S). The symbols at the end of selected rows illustrate the models included in Fig. 7.

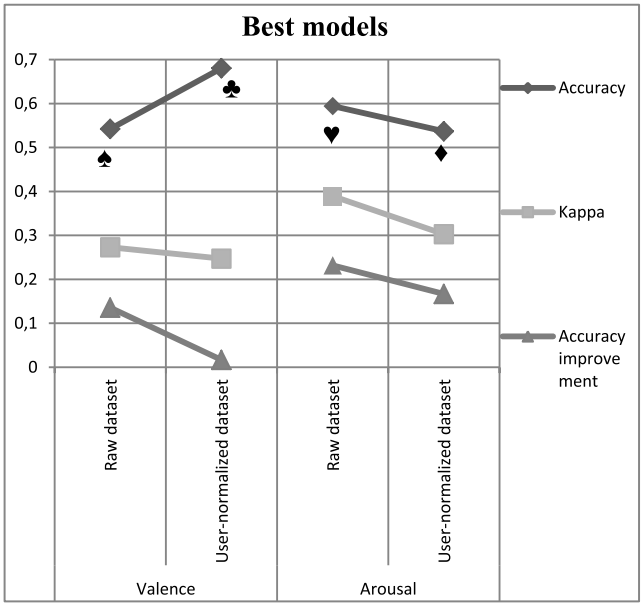

FIGURE 7. Models with highest Kappa value and their corresponding accuracy and accuracy improvement (i.e., the difference between model accuracy and base rate classifier accuracy) across different datasets and emotional dimensions.

As shown in the two result tables (i.e., Table 4 and Table 5), the user-normalized dataset reaches higher accuracy values. This can be due to the additional information features provide. Instead of raw interaction information, these features focus on participants' variations in their own interaction

TABLE 4. Models generated from the user-normalized dataset with best accuracy, kappa and classifier accuracy improvement (compared to the base rate classifier).

Class Attribute	Class Att. Values	Accuracy	Kappa	Ref. Acc.	Acc. Impr.	Algorithm	Settings
Arousal	Neg.(<1) non-neg. (≥1)	0.644	**0.285**	0.533	**0.111**	RF	D,F,P
		0.806	0.216	0.798	0.008	RF	S
	Pos. (>1) neut. (1) neg. (<1)	0.537	**0.303**	0.370	**0.166**	SVM	D,S ♦
		0.647	0.000	0.647	0.000	SVM	P
Valence	Neg.(<1) non-neg. (≥1)	0.619	0.235	0.523	**0.095**	RF	D,F,P
		0.680	**0.247**	0.663	0.016	RF	F,S,P ♣
	Pos. (>1) neut. (1) neg. (<1)	0.579	**0.223**	0.066	-0.084	NB	S
		0.663	0.0	0.663	0.0	SVM	P
		0.411	0.089	0.382	**0.029**	NB	D,F,S

TABLE 5. Models generated from the Raw dataset with best accuracy, kappa and classifier accuracy improvement (compared to the base rate classifier).

Class attribute	Class att. values	Accuracy	Kappa	Ref. Acc.	Acc. Impr.	Algorithm	Settings
Arousal	Pos. (>5) neut. (5) neg. (<5)	**0.483**	**0.196**	0.4	**0.083**	SVM	S
	pos. (>6) neut. (>4 and <6) neg. (<4)	**0.594**	**0.389**	0.362	**0.231**	SVM	D ♥
Valence	pos. (>5) neut. (5) neg. (<5)	0.489	0.230	0.347	**0.141**	RF	F,D
		0.516	**0.255**	0.4	0.116	SVM	
	pos. (>6) neut. (>4 and <6) neg. (<4)	0.542	**0.272**	0.406	**0.135**	RF	D,P ♠
		0.625	0.219	0.608	0.016	RF	S

behavior. Kappa values are also not very high (although those values can also be found in related works as can be seen in next paragraph). This can likely be due to the relatively small dataset (data from 41 participants extracted in real-world conditions) and a high number of features that eventually had to be reduced. Another salient feature is that the best kappa values and the best accuracy improvements (compared to the majority class classifier) were achieved in datasets where equal size sampling was applied. User-normalized dataset (Table 4) offers the best accuracy rates, providing similar results in arousal and valence (but slightly higher in 2-value classification than in 3-value classification). In the raw dataset (Table 5) lower accuracy values were obtained, with similar kappa and accuracy improvement values.

Comparing our results to similar work, we can find low kappa values from best-performing models in previous studies. For instance, in [16] the best model distinguishing engagement from boredom achieved a 0.374 kappa value (87.0% accuracy rate) and the best model distinguishing three emotions achieved a 0.171 kappa value (56.3% accuracy rate). These results are drawn from a study that has similar characteristics to our current work. In [16], researchers use data recorded from a 3-essay writing experiment (with about 1440 data instances, as they generate a data instance every 15 seconds), although their study did not use the mouse as a data source. Following the same goals for experimentation, other studies showed mixed results. Top classifiers from [5], predicting single emotions, provide accuracies from 76.3% (Kappa=0.55) to 93.8% (Kappa=0.55). The models from [5] were built using a reduced data set (initially reduced through the removal of a neutral category undersampling). Also, these results [5] were generated from fixed text models, as the free text models did not achieve kappa values over 0.4. Accuracy values presented in [9] varied depending on the emotion predicted and the algorithms used, achieving values from 47.37% to 81.25% (mean=62.47%, std. dev.=8.67). In the case of mouse-based detection, in [25] accuracy rates ranged from 91.96% to 94.61% when predicting confusion and data about a known target. When no information about the target is used, the accuracy rates go from 82.38% to 84.47%.

In order to evaluate the different methodological questions introduced in this paper, the highest kappa value for each variable has been selected and depicted in Figs. 7, 8, 9 and 10. Kappa statistic indicates how much the classification rate was a true reflection of the model or how much it could be attributed to chance alone [5]. Kappa has been used as an indicator to choose models that perform better than chance. The following models are depicted by the accuracy of the model, the accuracy improvement compared to the majority class (i.e. the difference between the accuracy reported by the model and the proportion of instances of the majority class) and the Cohen's Kappa. These models depicted in Fig. 7 are

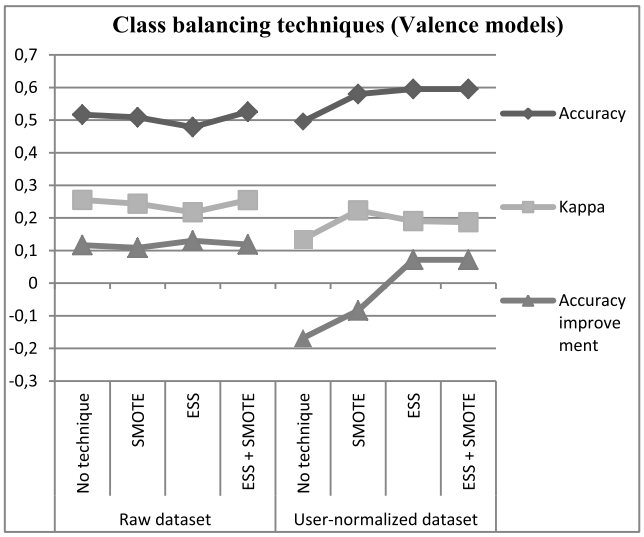

FIGURE 8. Valence models with highest kappa value and their corresponding accuracy and accuracy improvement (i.e., the difference between model accuracy and base rate classifier accuracy) across different class balancing techniques (*SMOTE* oversampling, Equal Size Sampling "*ESS*" and SMOTE oversampling applied after Equal Size Sampling "*ESS + SMOTE*") and datasets.

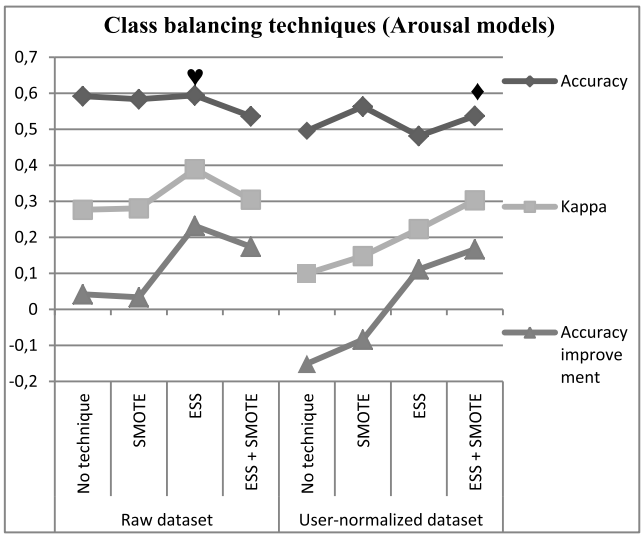

FIGURE 9. Arousal models with highest kappa value and their corresponding accuracy and accuracy improvement (i.e., the difference between model accuracy and base rate classifier accuracy) across different class balancing techniques (*SMOTE* oversampling, Equal Size Sampling "*ESS*" and SMOTE oversampling applied after Equal Size Sampling "*ESS + SMOTE*") and datasets.

identified in the following figures by the symbols ♠, ♣, ♡ and ♦.

Results in Fig. 7 show the impact of the different datasets when analyzing the models with the highest kappa value from the raw and the user-normalized dataset. The raw dataset provides similar (when predicting valence) or slightly better (when predicting arousal) Kappa values than the user-normalized version. User-normalized version model of valence shows 15% higher accuracy than the model based on the raw dataset. The finding that the user-normalized

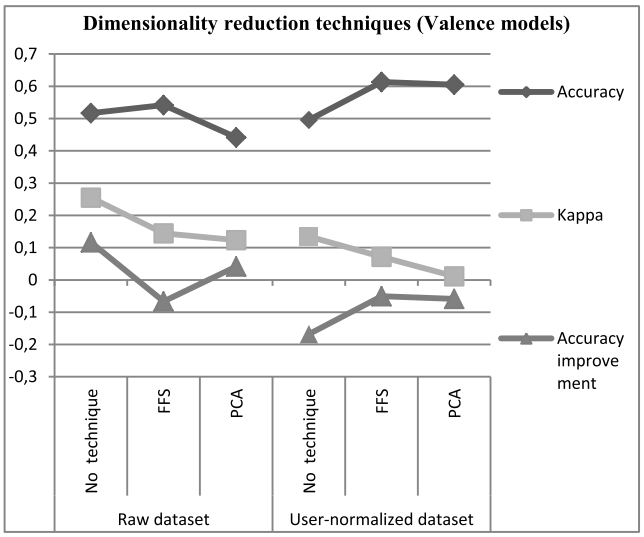

FIGURE 10. Valence models with highest kappa value and their corresponding accuracy and accuracy improvement (i.e., the difference between model accuracy and base rate classifier accuracy) across different dimensionality reduction techniques (Forward Feature Selection "FFS" and Principal Component Analysis "PCA") and datasets.

dataset performed worse on the accuracy improvement metric is likely due to the imbalanced class attribute (see Fig. 8 and Fig. 9). In conclusion, models seem to perform better when detecting the arousal dimension and when predicting arousal, the raw dataset seems to provide slightly better results than the user-normalized dataset.

To evaluate the impact of the different class balancing techniques, we assessed model performance based on the accuracy rates, kappa values and generated models with the highest kappa from each class balancing technique and dataset used (see Fig. 8 for valence and Fig. 9 for arousal). We can see how the raw dataset and the user-normalized dataset based models provide different results depending on the techniques used. While the application of class balancing techniques show almost no improvement on the models generated from the raw dataset (with an exception when applying ESS to predict arousal), these techniques seem to provide better results (compared to the best model with no technique) when used on the user-normalized dataset. All kappa and accuracy improvement values are higher when class balancing technique is used. Regarding techniques, SMOTE only provides better results (in comparison to the highest kappa model with no class balancing technique) when used in the user-normalized dataset. ESS surpasses SMOTE in most cases using the user normalized dataset, additionally, providing better or similar accuracy improvements and kappa values. Top accuracies obtained are around 60% with kappa values below 0.4. With these results, we can see how the use of class balancing techniques results in the user-normalized based models performing just as well as the raw dataset based models.

Fig. 10 and Fig. 11 show the impact of the different dimensionality reduction techniques with the different datasets with

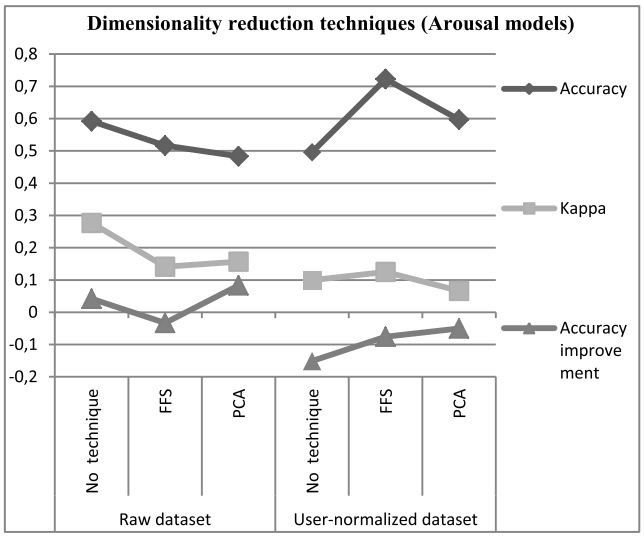

FIGURE 11. Arousal models with highest kappa value and their corresponding accuracy and accuracy improvement (i.e., the difference between model accuracy and base rate classifier accuracy) across different dimensionality reduction techniques (Forward Feature Selection "FFS" and Principal Component Analysis "PCA") and datasets.

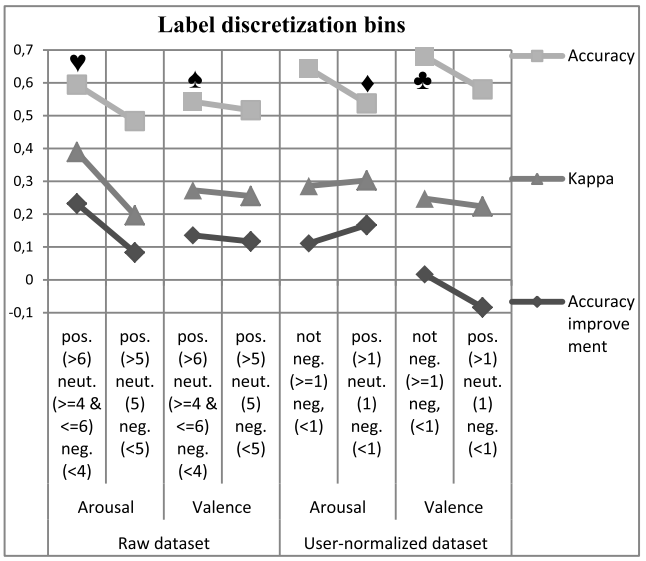

FIGURE 12. Models with highest Kappa value and their corresponding accuracy and accuracy improvement (i.e., the difference between model accuracy and base rate classifier accuracy) across different discretization bins applied to the affective label attribute and datasets.

the highest kappa value. As shown in these figures, the application of the proposed techniques provides a similar or worse kappa value of the models generated. In the raw dataset, the accuracy values are similar or lower, with a big decrease in accuracy improvement when using forward feature selection. When using the user-normalized dataset, none of the resultant models provide an accuracy improvement over the majority class and the kappa values. These models also performed lower than the ones generated from the raw dataset.

Some models from Fig. 7 do not appear in Figs. 8, 9, 10 or 11. This is because these models were generated by combining class balancing (Figs. 8 and 9) and dimensionality reduction techniques (Figs. 10 and 11). As these figures aim to study the impact of these techniques independently, those models are not shown there.

When analyzing the results regarding the different discretization thresholds used with the affective label attribute, it can be seen in Fig. 12 that approaches that contain a category with small range (i.e., the difference between the maximum and minimum values included in that category is smaller than the same difference of other categories) provide lower accuracy rates. This happens when the neutral bin contains only one value (5) instead of three (4, 5 and 6) in the raw dataset and when there is a small bin containing one value (neutral=1) in the user-normalized dataset. When taking a look at the kappa and the accuracy improvement, most of the models in Fig. 12 show a similar behavior than the one shown by accuracy, yielding higher values when there is no category with small amplitude compared to others. The only exception is the arousal model generated from the user normalized dataset. This model provides kappa values that are similar to and slightly better than the kappa and accuracy improvement values with a small amplitude category.

The overall values provided by the models is potentially the result of the difficulty of performing highly detailed predictions from labels grouped in unbalanced categories (containing fewer data instances, which may impact on the robustness of the model).

When evaluating the models with the highest kappa values across algorithms, similar results were found in the raw dataset models. Support vector machines achieved the best results followed by random forest for both valence and arousal. Naïve Bayes and C4.5 best models show worse values than support vector machines and random forest.

The models generated from the user-normalized dataset show similar kappa values when using support vector machines, random forest or Naïve Bayes (with better random forest showing best accuracies than the other algorithms). C4.5 best model show worse values than the others when generating the arousal model from the user normalized dataset.

Note that related work provides low values in the same range than the ones we have obtained [5], [16]. In particular, when there is a small number of data instances, a large number of features and the target outcome is subjective, it is very difficult to obtain better results. Nonetheless, our contribution has improvements over previous approaches. In particular, our work uses a discretized dimensional approach, while most similar studies use a closed set of emotions, which can produce different results depending on the emotion predicted [5], [9]. Another methodological aspect to consider is the impact of the proposed task on the results. In our case, we are dealing with a free text task, and this kind of task relates to low kappa values (in [5], results based on free text provide kappa values below 0.4).

Regarding all the variables that have been evaluated in this work, the following conclusions can be drawn from the results

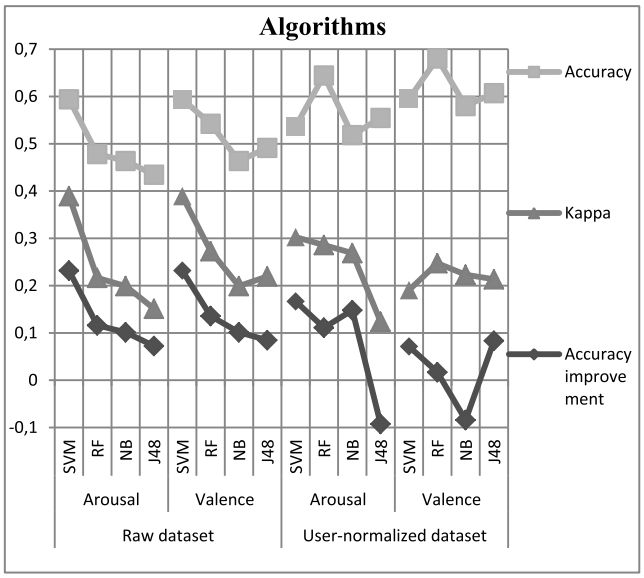

FIGURE 13. Models with highest Kappa value and their corresponding accuracy and accuracy improvement (i.e., the difference between model accuracy and base rate classifier accuracy) across different algorithms (Support Vector Machine "SVM", Random Forest "RF", Naïve Bayes "NB" and C4.5 tree "J48").

reported above: i) the use of a baseline in order to calculate user-normalized (i.e., using individual interaction patterns as referential values); mouse and keyboard features does not seem to offer better results compared to the generation of the raw mouse and keyboard features without using a baseline; ii) when using a discretization method with the class attribute, approaches with an unbalanced bin (e.g., a small neutral bin with positive and negative bins representing a wider range from the original variable spectrum) tend to offer worse accuracy rates; iii) the only data preprocessing technique whose best model provided better results were the class balancing ones (SMOTE and equal size sampling) on the user-normalized dataset, while dimensionality reduction techniques (forward feature selection and principal component analysis) did not provide significantly better results; and iv) support vector machines and random forests are the algorithms that achieved the best models.

V. DISCUSSION

We have identified and further explored the main modeling issues involved in unobtrusively detecting the affective state of a learner from keyboard and mouse interactions in essay writing tasks. We have automatically created multiple models (over 500) using data mining techniques where we have combined features and data sources proposed in different research works and extended them in order to address a real-world learning scenario.

A. RESULTS

Results displayed in Fig. 7 do not show an improvement in the models generated from the user-normalized dataset. Our evaluation of the different preprocessing techniques suggests that class balancing techniques used produced better results, than not using them, or similar ones, while dimensionality reduction did not produce better models. Applying class balancing techniques on the user-normalized dataset (Fig. 8 and Fig. 9) led to better results, achieving similar results to the models generated from the raw dataset models. This suggests the user-normalization may lead to unbalanced classes. It then makes sense to take a deeper look at two different points in our process that might have led to an unbalanced dataset: i) the normalization process (where, in case a very high or low value of the class attribute in the initial baseline might bias the class attribute of all the data instances of that subject) and ii) the class attribute discretization process. With respect to the former, the study of other approaches to calculate the reference value of the normalization could be explored (e.g. using a dynamic baseline, or using a dynamic window from the recent interactions as reference values).. With respect to the latter, regarding the different ways to discretize affective labeling, approaches that include categories with unbalanced ranges seem to yield lower accuracy rates (see Fig. 12).

B. BASELINE MODEL

As previously stated, the main hypothesis we wanted to test is that ''creating a baseline model to be used as a reference of how the participant interacts with the keyboard and mouse improves affect detection results''. We created an interaction baseline model that can be used as a reference for how the participant generally interacts with the keyboard and mouse. This way we provide new modeling alternatives which are based on leveraging a user's specific changes and comparing their values across different tasks with respect to the calibration task (i.e., where the baseline model is obtained).

The interaction baseline model's goal is to model keystroke-level features in relation to affect, thus enriching predictive models. This approach follows what is commonly applied when processing physiological signals [58]. Using baseline measurements as a reference model is not uncommon [2], [6], but our proposal differs in various ways. We have used a single baseline model to compare the user's behavior over several different tasks in a real-world scenario characterized by the shortage of data. As it can be seen in Fig. 7, the use of this interaction baseline can provide better accuracy rates than the use of raw interaction features when predicting affective states.

The calibration task to get the baseline model takes little time and is done once at the beginning of the experiment, thus fulfilling another requirement of our approach, which is to provide experimentation settings that closely represent natural learning settings and can be applied in real-world learning scenarios. This way we are trying to minimize the usage of a fixed text, not forcing the user to type it several times, as others have done to collect observable features directly from it [5], [7]. This has been an initial proposal that can be adjusted to different experimental conditions. For instance, a keyboard baseline model task could be repeated in a long-term experiment to study the validity of the proposed model

over time. The fixed text used in that model can change in order to recalculate baseline features and evaluate their usage. Another possibility in a long-term experiment is to replace this baseline model aggregating new features calculated from a very long time window, such as features from the same digraph typed many times during one day in a free-text data collection approach.

Regarding the choice of the text selected for the calibration task, related work used a text from a given popular book, with no apparent reason provided [7]. We took a similar text as in [5] because it provides relatively simple sentence structure with no long uncommon words, and each piece of text has roughly the same length [59]. Besides, it is appropriate for essay writing tasks, which are commonly performed by learners of English as second language, the target population in our experiment. The complexity of the proposed text has to be moderate in this case, as copying a text that the participant cannot understand may lead to confusion or frustration. All this raises another issue: the choice of the text may also impact the quality of the initial model. On the one hand, the shorter the fixed text is, the less intrusive it is for the participant (as the time the participant spends in that calibration task is time that is not being used on a ''real'' task). On the other hand, the longer the text and the more diverse words included, the more digraphs are modeled (so the model generated will be richer) [60]. So for a given population and educational context a balance has to be found between these two factors. The meaning of the text is another point to address, as it can elicit emotions by itself [61] so a neutral content text has to be selected.

C. EXPERIMENTAL ENVIRONMENT-RELATED LIMITATIONS

In this work we are building a model of a person in a real-world non-intrusive educational setting from keyboard and mouse interactions. There is related evidence showing that models which focus on individual person features tend to be more accurate, but in these settings there is lack of large interaction data sets from which to get an accurate model of the learner features, which is a well-known challenge [9]. This issue has driven us to explore additional modeling features based on different types of measurements. The problem here is that we are not dealing with person dependent models recorded in ideal conditions (usually obtained in non-authentic contexts) but with more naturalistic contexts where lower accuracies are obtained [2], [10]. Related work has shown that individual models are difficult to build because of difficulties in getting enough samples per user, and usually, those datasets in which predictions do not surpass certain accuracy thresholds are neglected [5], [9], [10].

Taking into account the relatively small number of instances available in the dataset, the overall results encourage us to continue to do research in this direction. The kappa values obtained are low due to the nature of the experimentation data, which is relatively more representative of a real-world scenario than previous fixed-text data. Hence, further research in this area with bigger datasets could help create more robust models. In this sense, a more robust baseline constructed from more data instances could help us reduce the noise in the data.

Bearing this dimensionality challenge in mind, from the modeling viewpoint, we have included combinations of the most frequent two and three keystroke events (i.e., digraphs and trigraphs) and combined these key-specific features with other key-independent features in order to detect affect state changes. Considering the small amount of instances available in our current dataset, the overall results suggest further research in this direction. Note that in classifier design, some papers suggest that there should be 10 times more instances (training samples) per class than the number of features [62], which are challenging given the circumstances of our setting (which are similar to others [6]). A long term version of the proposed experiment would enable us to extend the approach by providing more exercises over time and thus increase the number of instances in our datasets. Furthermore, instead of dealing with groups of students, focusing on personalized modeling, each participant may be involved in an within-subject study, which is expected to increase accuracy results [9]. This within-subject design would provide a more robust baseline and most likely additional performance features for each research subject, thus enabling to model sequences of actions more accurately and create more complex performance-based models [63].

In any case, the proposed approach in this work provides new modeling opportunities, evaluated in real-world scenarios with multiple users, which can be further explored in future experiments where more interactions could be involved. In particular, we plan to conduct a long-term version of the proposed experiment, using a within-subject approach, to study individual features over different real-world problems. This way, with more interaction data from each learner, we expect to provide a more robust baseline model and represent a wider range of the student's performance across features, to model sequences of actions and create more complex performance-based models [63].

D. DATA PREPARATION

Another key subject is the impact of using a particular data preprocessing method. One of the goals of this preprocessing is the high dimensionality of data in this field, which is a relatively common problem identified in literature [5], [27], [36]. This is usually tackled by means of different preprocessing and dimensionality reduction techniques. There are many dimensionality reduction techniques and their use can have an impact on different aspects of the model generated, ranging from model interpretation (e.g., PCA generated features are calculated by combining the original ones. As such it will not be possible to evaluate which set of original features have the most impact on the results) to the performance of the model generation, as some techniques, like forward feature selection, can be time consuming. In case of class balancing techniques, the use of undersampling based techniques is debatable when there are

very few instances in the dataset. In this work some of the most common preprocessing methods have been applied as a variable to be considered when generating the models. The purpose here is to evaluate their impact on the models generated, but usually related work uses them without evaluating their appropriateness. In our case (see Fig. 8), it is shown that some techniques (in our case the dimensionality reduction ones) may not result in an improvement of results. Additionally in some cases, such as forward feature selection, the suitability of the technique may still be brought into question because of the time consumption related to it. Regarding class balancing techniques, we have seen that the models from the user-normalized dataset improved their results when these techniques were used, which lead us to think that the user-normalized dataset could be unbalanced. This class unbalancing from the user-normalization process should be fixed in further experiments. To do that, differences in the user-normalized dataset generation should be carefully studied, from the data normalization to the label discretization. Although they have not been addressed in the work reported here, other technical issues related to the different preprocessing methods should be taken into account, such as time consumption. Due to the small size of our dataset, time consumption has not been a problem, but in large datasets, it could make a difference, particularly when the models proposed are used in other real world scenarios [64].

Another important aspect of the data preparation is providing the proposed approach with the capability of adapting to different experimentation conditions. By using appropriate data preprocessing techniques, the system might be adapted to tasks where one of the proposed data sources is not being used. For instance, this approach can be employed in educational games where only a mouse is needed [65], thus dealing with having a lack of information from other data sources [36]. This approach can also be explored to adapt affect detection to people with special needs. In this case, [42] suggests that some of the proposed data sources used in this work may change their interaction purposes. For instance, blind people use a keyboard for navigating over the materials with a screen reader, and we could detect and compare their keystroke behavior when they are either navigating or writing [42].

To tackle this issue, different classifiers could be used for each data source and the different models generated could be combined, thus determining which data source could offer more information for each user in a given situation over time. The use of unsupervised learning techniques can also be integrated with the research described here, thus generating groups of similar users and generating models for the users depending on their group, following related approaches in exploratory learning environments [66]

E. AFFECTIVE STATE REPRESENTATION

Another variable addressed in the current work is which affective model representation to use [67]. As previously introduced (see Section I), we have adopted a simplified dimensional approach for labeling the user's affective state, which is readily available and can be managed by students themselves [12]. The problem of labeling affect is well known, and here inter-observer agreement can be low [67]. The use of sets of emotions, where the participant has to choose the closest to her current state, is common in the literature. This categorical approach can result in different interpretations of the same emotion [6]. We have dealt with the potential problem of affective concept misinterpretation by providing learners with a written explanation of the two affect dimensions covered in this approach, valence and arousal, in order to ensure that all learners understand the dimensions in the same way. Other approaches deal with this issue by providing mappings from affective dimensions to affective categories, even when they are focused on learning-centered cognitive-affective states [68]. Although this categorical approach is out of our scope for the reasons discussed above, there is evidence that coders who have received proper thorough methodological training assessing affect have achieved inter-rate reliability rates of over 0.6 [69].

There are other ways of detecting participants' affective state, such as using a free text approach [70], which we have explored in the past [43]. This approach has not been followed here since we consider ordinary daily practices of learners, and we do not want to distract them with additional affect-oriented tasks.

Our work has explored different ways of discretizing the dimensional values obtained into different dimensional categories to perform the prediction. This approach seems to be interesting since it provides a finer-grained means of modeling students' affective states. However, this would hamper the simplicity we are aiming at. We have selected a discretized approach that simplifies the problem of predicting the dimensional values of the affective state for the algorithms used. By using positive, negative or neutral states (from the raw dataset) or positive or negative emotional transitions (from the user-normalized dataset) we are keeping a reduced number of categories, which are easier to manage when there is a shortage of data. In this sense, more discretization approaches could have been evaluated, thus following this paper's approach, which shows the first step at looking for a proposal that supports classification algorithms and also is able to provide a balance between meaningfulness and simplicity for users.

The emotional modeling approach deserves particular attention when the affective state detector is to be integrated with other components that will use the predictions performed [71]. This is another reason why this dimensional approach was chosen, as it provides a more flexible and standardized description to select which affective state phenomena to take into account. From the results obtained, we can see that the 2-value classification also offers slightly better results than the 3-value approach. Although some previous works use high/low or positive/negative values for the proposed dimensions [32], [33], [47], [72], the use of neutral states could also be taken into account, for example, to use

a two-step classification approach able to discard those neutral states before classifying positive or negative affective states [40].

Regarding the evaluation of the work done in the discretization approaches for the labeling attribute, further research could be performed, in which a recommender system employs the different approaches evaluated. A recommender system would give us a closer look at the impact of discretization criteria on the recommendations given, towards understanding what granularity is most important in different situations. The characterization of affect determines the type of recommendations that can be implemented in real-world scenarios. Recommendations could be provided in terms of well-known traditional interaction sources, such as the ones used here (i.e., mouse and keyboard) or less explored interaction sources, such as visual, sound or haptic, which can be tuned to provide recommendations in ambient intelligent scenarios [73], [74].

F. DATA SOURCES

Another key issue in this work is the suitability of the interaction devices that are used and modeled. Although the task in this study mainly depends on typing, the mouse was used by all the participants for navigation and for text editing purposes. In a series of tasks where the keyboard is rarely used, automatic feature selection methods should discard the keyboard features. Similarly, a judicious selection of threshold values (like the ones depicted in Fig. 5), also ensures that features are only used when they represent an interaction repeated a potentially significant number of times.

As to the model generation and the device used, although all the participants interacted through keyboard and mouse in this work, it is common nowadays to find users controlling desktop computers (or other devices) by other means, such as track pads or touchscreens. The link between touchscreen based typing and mood changes was investigated in [75], which uses digraphs and trigraphs and takes into account other touchscreen related features, such as the number of hands used for typing. In [76], the use of a touchscreen is also combined with keyboard and mouse features for affective computing. Indeed, there is a vast literature evaluating different data sources as data for affective state detectors [34]. Other point to discuss is the possible combination or impact of using the proposed data sources with other data sources (e.g. using psychophysiological data sources in touch typing tasks [77]). Here, we tried to stick to a low-intrusiveness and cheap approach that requires no additional hardware in order to evaluate this work in real-world conditions.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we implemented a case study which shows the critical issues involved in detecting affect from traditionally non-intrusive interaction sources, such as keyboard and mouse. The approach proposed aims at deepening the information that can be gathered and modeled from these interaction sources so that these are fully optimized in learning scenarios.

With our focus on real-world scenarios, we took intrusiveness into account in the experimental design by not distracting the user with affect-oriented tasks while performing learning tasks. The intrusiveness has also been taken into account in the choice of the proposed data sources, as this work is based on two of the most basic interaction sources to date, keyword and mouse. Through the user study reported in this paper we have shown that these two interaction sources can be further explored to better account for the learners' affect. Ultimately, our approach aims to clarify and take advantage of the modeling issues involved in detecting affect from diverse users' interaction sources. Although the choice of interaction sources depends on the given goals and educational context, traditional input sources are currently playing an important role.

Our work has investigated three main topics: i) creating an interaction baseline model for each user from which to model user features related to several dimensions, ii) diminishing the dimensionality of mouse and keyboard features and balance the training instances by taking preprocessing techniques as a methodological variable, and iii) adopting a simplified characterization approach for labeling the users' affective state.

Following the open issues discussed in the previous section, future work could be applied to the following directions: 1) explore different approaches to calculate the reference value used within normalization, 2) evaluate discretization approaches for the class attribute, 3) explore new approaches for the baseline proposed, such as a dynamic baseline, 4) conduct a long-term within-subject experiment to study individual features' applicability within different real-world problems, 5) explore other interaction devices available in mobile devices, 6) analyze the impact of discretization labeling when using a recommender system, 7) evaluate the impact of personality or demographic features (such as gender or age) on the generated models. There are many possibilities for future work in this rapidly-developing field.

In conclusion, this research has investigated some open modeling issues found in related literature and also has proposed a new approach aiming to highlight user personal variations in interaction data (from mouse and keyboard) collected in an between-subject experiment in order to predict affective states. We hope that the results presented in this paper encourage further research in finding suitable approaches to deal with the problem of detecting affect in realistic educational scenarios, in which collected data are sparse and imbalanced.

ACKNOWLEDGMENTS

Authors would like to thank Maria del Mar Saneiro and Raul Uria (members of aDeNu Research group) for their work in the design and running of the experiments, as well as Ma. Victoria Almeda for her help and tips on this paper and the rest members of the Baker EDM lab at Teachers College, Columbia University for their support.

REFERENCES

[1] D. Putwain, P. Sander, and D. Larkin, "Academic self-efficacy in study-related skills and behaviours: Relations with learning-related emotions and academic success," *Brit. J. Educ. Psychol.*, vol. 83, no. 4, pp. 633–650, 2013.

[2] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 43:1–43:36, Feb. 2015.

[3] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[4] D. Novak, M. Mihelj, and M. Munih, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interact. Comput.*, vol. 24, no. 3, pp. 154–172, May 2012.

[5] C. Epp, M. Lippold, and R. L. Mandryk, "Identifying emotional states using keystroke dynamics," in *Proc. Annu. Conf. Human Factors Comput. Syst.*, May 2011, pp. 715–724.

[6] A. Kołakowska, "A review of emotion recognition methods based on keystroke dynamics and mouse movements," in *Proc. 6th Int. Conf. Human Syst. Interact. (HSI)*, Jun. 2013, pp. 548–555.

[7] P. Shukla and R. Solanki, "Web based keystroke dynamics application for identifying emotional state," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 11, pp. 4489–4493, Nov. 2013.

[8] R. Shikder, S. Rahaman, F. Afroze, and A. B. M. A. Al Islam, "Keystroke/mouse usage based emotion detection and user identification," in *Proc. Int. Conf. Netw., Syst. Secur. (NSysS)*, Jan. 2017, pp. 96–104.

[9] A. Kołakowska, "Recognizing emotions on the basis of keystroke dynamics," in *Proc. 8th Int. Conf. Human Syst. Interact. (HSI)*, Jun. 2015, pp. 291–297.

[10] D. Gunetti and C. Picardi, "Keystroke analysis of free text," *ACM Trans. Inf. Syst. Secur.*, vol. 8, no. 3, pp. 312–347, Aug. 2005.

[11] X. Li *et al.*, "Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information," *PLoS Biol.*, vol. 15, no. 1, p. e2001402, Jan. 2017.

[12] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Therapy Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[13] C. Romero and S. Ventura, "Educational data science in massive open online courses," *Wiley Interdiscipl. Rev. Data Mining Knowl. Discovery*, vol. 7, no. 1, p. e1187, Jan. 2017.

[14] H. Hamdi, P. Richard, A. Suteau, and P. Allain, "Emotion assessment for affective computing based on physiological responses," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jun. 2012, pp. 1–8.

[15] Y. Zhong and Y. Deng, "A survey on keystroke dynamics biometrics: Approaches, advances, and evaluations," in *Recent Advances in User Authentication Using Keystroke Dynamics Biometrics*. Science Gate Publishing, 2015, pp. 1–22.

[16] R. Bixler and S. D'Mello, "Towards automated detection and regulation of affective states during academic writing," in *Artificial Intelligence in Education*, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Berlin, Germany: Springer, 2013, pp. 904–907.

[17] W.-H. Tsui, P. Lee, and T.-C. Hsiao, "The effect of emotion on keystroke: An experimental study using facial feedback hypothesis," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 2870–2873.

[18] D. G. Brizan, A. Goodkind, P. Koch, K. Balagani, V. V. Phoha, and A. Rosenberg, "Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics," *Int. J. Hum.-Comput. Stud.*, vol. 82, pp. 57–68, Oct. 2015.

[19] P.-M. Lee, W.-H. Tsui, and T.-C. Hsiao, "The influence of emotion on keyboard typing: an experimental study using visual stimuli," *Biomed. Eng. OnLine*, vol. 13, no. 1, p. 81, Jun. 2014.

[20] Y. M. Lim, A. Ayesh, and M. Stacey, "Detecting cognitive stress from keyboard and mouse dynamics during mental arithmetic," in *Proc. Sci. Inf. Conf. (SAI)*, Aug. 2014, pp. 146–152.

[21] S. Salmeron-Majadas, O. C. Santos, and J. G. Boticario, "An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context," in *Proc. 18th Annu. Conf. Knowl.-Based Intell. Inf. Eng. Syst. (KES)*, Gdynia, Poland, vol. 35. 2014, pp. 691–700.

[22] A. F. M. N. H. Nahin, J. M. Alam, H. Mahmud, and K. Hasan, "Identifying emotion by keystroke dynamics and text pattern analysis," *Behav. Inf. Technol.*, vol. 33, no. 9, pp. 987–996, Sep. 2014.

[23] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Technical manual and affective ratings," Center Res. Psychophysiol., Univ. Florida, Gainesville, FL, USA, Tech. Rep., 1999.

[24] M. T. Hibbeln, J. L. Jenkins, C. Schneider, J. Valacich, and M. Weinmann, "Inferring negative emotion from mouse cursor movements," Social Sci. Res. Netw., Rochester, NY, USA, SSRN Scholarly Paper ID 2708108, 2016.

[25] A. Pentel, "Patterns of confusion: Using mouse logs to predict user's emotional state," in *Proc. Int. Workshop Pers. Approaches Learn. Environ.. (PALE)*, 2015, pp. 40–45.

[26] P. Lali, M. Naghizadeh, H. Nasrollahi, H. Moradi, and M. S. Mirian, "Your mouse can tell about your emotions," in *Proc. 4th Int. eConf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2014, pp. 47–51.

[27] T. Yamauchi, "Mouse trajectories and state anxiety: Feature selection with random forest," in *Proc. Humaine Association Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 399–404.

[28] L. Vea and M. M. Rodrigo, "Modeling negative affect detector of novice programming students using keyboard dynamics and mouse behavior," in *Proc. Trends Artif. Intell., PRICAI Workshops*, 2016, pp. 127–138.

[29] J. Deng, S. Frühholz, Z. Zhang, and B. Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning," *IEEE Access*, vol. 5, pp. 5235–5246, 2017.

[30] B. Kabi, A. Samantaray, P. Patnaik, and A. Routray, "Voice cues, keyboard entry and mouse click for detection of affective and cognitive states: A case for use in technology-based pedagogy," in *Proc. IEEE 5th Int. Conf. Technol. Educ. (T4E)*, Dec. 2013, pp. 210–213.

[31] B. Lehman, M. Matthews, S. D'Mello, and N. Person, "What are you feeling? Investigating student affective states during expert human tutoring sessions," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2008, pp. 50–59.

[32] P. Zimmermann, S. Guttormsen, B. Danuser, and P. Gomez, "Affective computing—A rationale for measuring mood with mouse and keyboard," *Int. J. Occupation Safety Ergonom.*, vol. 9, no. 4, pp. 539–551, 2003.

[33] I. A. Khan, W.-P. Brinkman, and R. Hierons, "Towards estimating computer users' mood from interaction behaviour with keyboard and mouse," *Front. Comput. Sci.*, vol. 7, no. 6, pp. 943–954, Dec. 2013.

[34] O. C. Santos, "Emotions and personality in adaptive e-learning systems: An affective computing perspective," in *Emotions and Personality in Personalized Services*. Cham, Switzerland: Springer, 2016, pp. 263–285.

[35] K. Porayska-Pomsta, M. Mavrikis, S. D'Mello, C. Conati, and R. S. J. D. Baker, "Knowledge elicitation methods for affect modelling in education," *Int. J. Artif. Intell. Educ.*, vol. 22, no. 3, pp. 107–140, 2013.

[36] N. Bosch, H. Chen, S. D'Mello, R. Baker, and V. Shute, "Accuracy vs. availability heuristic in multimodal affect detection in the wild," in *Proc. ACM Int. Conf. Multimodal Interact.*, New York, NY, USA, 2015, pp. 267–274.

[37] N. Bosch *et al.*, "Automatic detection of learning-centered affective states in the wild," in *Proc. Int. Conf. Intell. Interfaces (IUI)*. New York, NY, USA: ACM, 2015, pp. 379–388.

[38] S. Craig, A. Graesser, J. Sullins, and B. Gholson, "Affect and learning: An exploratory look into the role of affect in learning with AutoTutor," *J. Educ. Media*, vol. 29, no. 3, pp. 241–250, 2004.

[39] C. Kim and R. Pekrun, "Emotions and motivation in learning and performance," in *Handbook of Research on Educational Communications and Technology*, J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop, Eds. New York, NY, USA: Springer, 2014, pp. 65–75.

[40] S. Salmeron-Majadas *et al.*, "Filtering of spontaneous and low intensity emotions in educational contexts," in *Artificial Intelligence in Education*, C. Conati, N. Heffernan, A. Mitrovic, and M. F. Verdejo, Eds. Madrid, Spain: Springer, 2015, pp. 429–438.

[41] S. Salmeron-Majadas, O. C. Santos, and J. G. Boticario, "Affective state detection in educational systems through mining multimodal data sources," in *Proc. 6th Int. Conf. Educ. Data Mining*, Memphis, TN, USA, 2013, pp. 348–349.

[42] O. C. Santos, A. Rodriguez-Ascaso, J. G. Boticario, S. Salmeron-Majadas, P. Quirós, and R. Cabestrero, "Challenges for inclusive affective detection in educational scenarios," in *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*, C. Stephanidis and M. Antona, Eds. Berlin, Germany: Springer, 2013, pp. 566–575.

[43] O. C. Santos, S. Salmeron-Majadas, and J. G. Boticario, "Emotions detection from math exercises by combining several data sources," in *Artificial Intelligence in Education*, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Berlin, Germany: Springer, 2013, pp. 742–745.
[44] M. M. Bradley and P. J. Lang, "The international affective digitized sounds (IADS-2): Affective ratings of sounds and instruction manual," Center Res. Psychophysiol., Univ. Florida, Gainesville, FL, USA, Tech. Rep. B-3, 2007.
[45] M. Lewis, S. M. Alessandri, and M. W. Sullivan, "Differences in shame and pride as a function of children's gender and task difficulty," *Child Develop.*, vol. 63, no. 3, pp. 630–638, Jun. 1992.
[46] J. B. Wiggins, J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "The relationship between task difficulty and emotion in online computer programming tutoring," in *Proc. 45th ACM Tech. Symp. Comput. Sci. Educ.*, New York, NY, USA, 2014, p. 721.
[47] P. G. Zimmermann, P. Gomez, B. Danuser, and S. Schär, "Extending usability: Putting affect into the user-experience," *Proc. NordiCHI*, 2006, pp. 27–32.
[48] O. C. Santos and J. G. Boticario, "Practical guidelines for designing and evaluating educationally oriented recommendations," *Comput. Educ.*, vol. 81, pp. 354–374, Feb. 2015.
[49] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 14th Int. Conf. Mach. Learn.*, 1997, pp. 179–186.
[50] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[51] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.
[52] I. Jolliffe, "Principal component analysis," in *Wiley StatsRef: Statistics Reference Online*. Hoboken, NJ, USA: Wiley, 2014.
[53] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 1993.
[54] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[55] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," in *Advances in Neural Information Processing Systems 10*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds. Cambridge, MA, USA: MIT Press, 1998, pp. 507–513.
[56] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *Proc. 11th Conf. Uncertainty Artif. Intell.*, San Francisco, CA, USA, 1995, pp. 338–345.
[57] M. R. Berthold *et al.*, "KNIME—The Konstanz information miner: Version 2.0 and beyond," *SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 26–31, Nov. 2009.
[58] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.
[59] C. Epp, "Identifying emotional states through keystroke dynamics," Dept. Comput. Sci., Univ. Saskatchewan, Saskatoon, SK, Canada, Tech. Rep., 2010.
[60] A. A. Ahmed and I. Traore, "Biometric recognition based on free-text keystroke dynamics," *IEEE Trans. Cybern.*, vol. 44, no. 4, pp. 458–472, Apr. 2014.
[61] M. Ainley, M. Corrigan, and N. Richardson, "Students, tasks and emotions: Identifying the contribution of emotions to students' reading of popular culture and popular science texts," *Learn. Instruct.*, vol. 15, no. 5, pp. 433–447, Oct. 2005.
[62] A. K. Jain, R. P. W. Duin, and J. C. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.
[63] J. M. L. Andres and M. M. T. Rodrigo, "Analyzing student action sequences and affect while playing physics playground," in *Proc. Int. Workshop Affect, Meta-Affect, Data Learn. (AMADL 2015)*, 2015, pp. 24–33.
[64] A. F. Botelho, R. S. Baker, and N. T. Heffernan, "Improving sensor-free affect detection using deep learning," in *Proc. Artif. Intell. Educ.*, 2017, pp. 40–51.
[65] J. L. Sabourin, L. R. Shores, B. W. Mott, and J. C. Lester, "Understanding and predicting student self-regulated learning strategies in game-based learning environments," *Int. J. Artif. Intell. Educ.*, vol. 23, nos. 1–4, pp. 94–114, Nov. 2013.
[66] L. Fratamico, C. Conati, S. Kardan, and I. Roll, "Applying a framework for student modeling in exploratory learning environments: Comparing data representation granularity to handle environment complexity," *Int. J. Artif. Intell. Educ.*, vol. 27, no. 2, pp. 320–352, Jun. 2017.
[67] S. K. D'Mello, "On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 136–149, Apr. 2016.
[68] R. S. J. D. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser, "Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments," *Int. J. Hum.-Comput. Stud.*, vol. 68, no. 4, pp. 223–241, Apr. 2010.
[69] J. Ocumpaugh, "Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual," Ateneo Lab. Learn. Sci., Columbia Univ., New York, NY, USA, Tech. Rep., 2015. [Online]. Available: http://www.upenn.edu/learninganalytics/ryanbaker/bromp.html
[70] S. E. Schultz *et al.*, "Blinded by science?: Exploring affective meaning in students' own words," in *Proc. Int. Conf. Intell. Tutoring Syst.*, 2016, pp. 314–319.
[71] I. Arroyo, B. P. Woolf, W. Burelson, K. Muldner, D. Rai, and M. Tai, "A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect," *Int. J. Artif. Intell. Educ.*, vol. 24, no. 4, pp. 387–426, Dec. 2014.
[72] D. A. M. Felipe, K. I. N. Gutierrez, E. C. M. Quiros, and L. A. Vea, "Towards the development of intelligent agent for novice C/C++ programmers through affective analysis of event logs," in *Proc. Int. MultiConf. Eng. Comput. Sci.*, vol. 1, pp. 511–518, 2012.
[73] O. C. Santos, R. Uria-Rivas, M. C. Rodriguez-Sanchez, and J. G. Boticario, "An open sensing and acting platform for context-aware affective support in ambient intelligent educational settings," *IEEE Sensors J.*, vol. 16, no. 10, pp. 3865–3874, May 2016.
[74] M. A. Eid and H. Al Osman, "Affective haptics: Current research and future directions," *IEEE Access*, vol. 4, pp. 26–40, 2016.
[75] A. Piscitello, "BiAffect: A system for analyzing neurocognitive functioning using keystroke dynamics and machine learning," M.S. thesis, Graduate College, Univ. Illinois Chicago, Chicago, IL, USA, 2015.
[76] K. Bakhtiyari, M. Taghavi, and H. Husain, "Hybrid affective computing—keyboard, mouse and touch screen: From review to experiment," *Neural Comput. Appl.*, vol. 26, no. 6, pp. 1277–1296, Aug. 2015.
[77] M. Meža, J. Košir, G. Strle, and A. Košir, "Towards automatic real-time estimation of observed learner's attention using psychophysiological and affective signals: The touch-typing study case," *IEEE Access*, vol. 5, pp. 27043–27060, 2017.

SERGIO SALMERON-MAJADAS received the degree in computer science at UCM in 2011 and the M.Sc. degree in advance artificial intelligence at UNED in 2014, with the eMadrid Network Special Mention to the Best Master's Degree Thesis. He is currently pursuing the Ph.D. degree in intelligent systems. He has been a member of the aDeNu Research Group, UNED, for four years. His research focuses on the use of machine learning techniques in order to predict affective states using different data sources.

RYAN S. BAKER is currently an Associate Professor with the University of Pennsylvania and the Director of the Penn Center for Learning Analytics. His lab conducts research on engagement and robust learning within online and blended learning, seeking to find actionable indicators that can be used today but which predict future student outcomes. He has developed models that can automatically detect student engagement in over a dozen online learning environments and has led the development of an observational protocol and app for field observation of student engagement that has been used by over 150 researchers in four countries. He has co-authored published papers with over 250 colleagues. He was the Founding President of the International Educational Data Mining Society. He is currently serving as an Associate Editor of two journals, has taught four MOOC instances, and was the first Technical Director of the Pittsburgh Science of Learning Center DataShop, the world's largest public repository for data on the interactions between learners and online learning environments.

OLGA C. SANTOS is currently an Assistant Professor (certified by Spanish ANECA up to the position of Associate Professor) with the Artificial Intelligence Department, UNED, Spain, and the Research and Development Technical Manager of the aDeNu Research Group. Her current research interests include combining artificial intelligence with ambient intelligence and Internet of Things to support personalized affective psychomotor learning that ubiquitously and dynamically adapt to the evolving user needs. She has participated in 16 research projects (UE, National), published over 150 papers, and co-chaired several workshop series (TUMAS-A, RecSysTEL/EdRecSys, PALE, and RSyL) and conferences (AIED, EDM, UMAP, and EC-TEL). She received the Best Doctoral Thesis Award by the IEEE Spanish Chapter of the Education Society and the 2014 Young Researcher Award of the IEEE Technical Committee on Learning Technology.

JESUS G. BOTICARIO has held several positions at UNED related to using IT in education. He is currently a Full Professor with the Computer Science School. He is the Head of the aDeNu Research Group, a Scientific Coordinator in european and national projects, the Chair of the CORE conferences, and an Organizer of the Workshop Series on User Modeling and Accessibility. He is also an Expert Counsellor at the UNED's Centre for Supporting Students with Disabilities and at the Corporate Social Responsibility Chair, UNED, and the Scientific Chairman of the Red ALTER-NATIVA. He has authored over 200 research articles and participated in 26 research and development funded projects.