Non-Acted Text and Keystrokes Database and Learning Methods to Recognize Emotions

MADIHA TAHIR, ZAHID HALIM, and ATTA UR RAHMAN, Machine Intelligence Research Group (MInG), Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology
MUHAMMAD WAQAS, Beijing Key Laboratory of Trusted Computing, Faculty of Information Technology, Beijing University of Technology
SHANSHAN TU, Faculty of Information Technology, Beijing University of Technology
SHENG CHEN, School of Electronics and Computer Science, University of Southampton
ZHU HAN, Department of Electrical and Computer Engineering, University of Houston

The modern computing applications are presently adapting to the convenient availability of huge and diverse data for making their pattern recognition methods smarter. Identification of dominant emotion solely based on the text data generated by humans is essential for the modern human–computer interaction. This work presents a multimodal text-keystrokes dataset and associated learning methods for the identification of human emotions hidden in small text. For this, a text-keystrokes data of 69 participants is collected in multiple scenarios. Stimuli are induced through videos in a controlled environment. After the stimuli induction, participants write their reviews about the given scenario in an unguided manner. Afterward, keystroke and in-text features are extracted from the dataset. These are used with an assortment of learning methods to identify emotion hidden in the short text. An accuracy of 86.95% is achieved by fusing text and keystroke features. Whereas, 100% accuracy is obtained for pleasure-displeasure classes of emotions using the fusion of keystroke/text features, tree-based feature selection method, and support vector machine classifier. The present work is also compared with four state-of-the-art techniques for the same task, where the results suggest that the present proposal performs better in terms of accuracy.

CCS Concepts: • **Applied computing** → *Word processors*;

Additional Keywords and Phrases: Affective computing, machine learning, affective states, pattern recognition, data-driven decision-making

The authors would like to thank the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, for providing research funding and facilities. This work was sponsored by the GIK Institute graduate research fund under GA-4 scheme. The authors also wish to thank the anonymous reviewers for spending time in reviewing and providing with important comments to improve the manuscript quality.

Authors' addresses: M. Tahir, Z. Halim (corresponding author), and A. U. Rahman, Machine Intelligence Research Group (MInG), Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, 23640, Topi, KP, Pakistan; emails: {madiha.tahir, zahid.halim, atta.rahman}@giki.edu.pk; M. Waqas and S. Tu, Beijing Key Laboratory of Trusted Computing, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, Beijing, China; emails: engr.waqas2079@gmail.com, sstu@bjut.edu.cn; S. Chen, School of Electronics and Computer Science, University of Southampton, SO17 1BJ, Southampton, United Kingdom; email: sqc@soton.ac.uk; Z. Han, Department of Electrical and Computer Engineering, University of Houston, 77204, Houston, USA; email: zhan2@central.uh.edu.

1551-6857/2022/02-ART61 $15.00
https://doi.org/10.1145/3480968

ACM Reference format:
Madiha Tahir, Zahid Halim, Atta Ur Rahman, Muhammad Waqas, Shanshan Tu, Sheng Chen, and Zhu Han. 2022. Non-Acted Text and Keystrokes Database and Learning Methods to Recognize Emotions. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 2, Article 61 (February 2022), 24 pages.
https://doi.org/10.1145/3480968

1 INTRODUCTION

The modern computing systems are presently facing the challenge of Big Data. The term Big Data is coined primarily to represent three Vs, i.e., the volume of the data, data velocity, and variety in the data [1]. In order to improve their services, modern enterprises need to employ data science methods to extract nontrivial information from the Big Data [2]. Identifying dominant human emotion is one such nontrivial task [3]. Emotion is a psycho and physiological process that influences lifestyle and interaction with the society. Human's actions, choices, and perception of an object or situation is influenced by various emotions [4]. According to Paul Ekman, a renowned psychologist, individuals with inborn facial paralysis have difficulty in developing and maintaining social relationships. Expressing emotions is important in the development of interpersonal communication [5]. The study of human emotional states and their influence has been a research area in the field of psychology and computer science. However, its application in the context of limited text has been overlooked in the past. The theories on affective computing are used in the development of human affective systems [6]. Paul Ekman identified six basic emotions in human beings belonging to any culture [5]. These include: happiness, sadness, anger, fear, disgusting, and surprise. Robert Plutchick, a psychologist, introduced the *wheel of emotions* that works on the principle of colors. According to that, emotion can be mixed to form different feelings, like joy and trust [7]. A psychologist, Silvan Tomkins's theory states that there are nine effects. He labeled them according to their intensity and physiological expressions. Tomkins categorized emotions into positive, neutral, and negative classes. In the negative class, anger, disgust, "dissmell" (a made-up word by Tomkins), distress, fear, and shame are included. According to him, the mental health of a person depends on the positive effects [8]. The current work utilizes Paul Ekman's theory on basic human emotions. This theory describes six emotions, namely, happy, angry, disgusting, fear, sad, and surprise [5]. To detect human effects, multiple methods have been proposed in the past [9–11] and a variety of approaches are used to manipulate human effects, like electroencephalography signals, speech signals, facial expressions, text data, and mouse/keystrokes' dynamics [12–15]. Emotions are the expressions made/measured through verbal ways, like speech, physical means, eye movements, **electroencephalogram (EEG)** signal, **electrocardiograph (ECG)** signal, and writing pattern.

Human emotion recognition during various communication instances over the digital media is an emerging area of research. Service providers are always interested to know about various emotions of their customers regarding the provided services/products. For example, customers are asked to select an emoji after getting a service. They have also requested to fill-in a brief feedback form and may also be invited to write a short review. The data gathered through this is used for the **Continual Quality Improvement (CQI)** purpose. Emotion identification using the **Computational Intelligence (CI)** methods in recent ages has become an active research problem because in almost every domain involving human interaction one needs to extract the feelings of persons interacting with the computing devices. For example, in the field of medicine, the detection of patients' emotion is important to give better treatment. Research on the detection of emotions from emails, tweets, short messages, and comments on YouTube are used to improve the service quality by the enterprises. At times, there are many emotions hidden in the short text collected from various consumers that can be of value to the enterprises. However, extracting this information from the text is a nontrivial task. This work utilizes the short text patterns of a person to detect her emotions using in-text data and the information associated with the keystrokes. Emotion recognition by the fusion of different modalities is a challenging problem. Various models are used to identify human emotions. Text and audio features are fused in [15] using a three-layer deep neural network, which considers spatial information from the text and temporal information from the audio data. Visual features from video and speech features are used in [16] for multimodal emotion detection. They apply the decision-level fusion technique to evaluate the probability of each modality and emotion. The works in [15, 16] develop DEAP and BAUM-I databases for multimodal emotion detection.

1.1 Human Emotion

Emotion is a biological state linked to the human nervous system. It brings in changes associated with one's feelings, thoughts, and attitude. Emotions play an important role in the daily life of human beings. An individual, directly or indirectly, expresses her emotion in almost every activity, e.g., while listening to music, handing over an item to another person, or while chatting with friends and relatives. Generally, there are two aspects of human emotions, the first is that there are no human experiences that are free of emotion, and the second is that all our memories of the past events preserve emotions that are experienced when those memories get recalled [17]. According to the domain of psychology, emotion is a complex state of feeling that results in physical and psychological changes that influence thoughts and behavior. Emotions can be expressed through verbal communication, gestures, actions, and written expression. In addition to the intentional expression of human emotion, at times, one tends to hide these as well. However, emotion hidden in a communication can be identified through various dominant cues. This can be conveniently identified in a face-to-face communication. However, for the communication that is made from a long distance, it is a challenging task. Such communication usually involves the text-based modes of message passing, like e-mails and letters. Owing to the recent advancements in the domains of information, communication, and technology, exchange of views on current trends/events irrespective of geographical borders has increased exponentially. This has been achieved through various online social networking websites (like, Twitter and Facebook) and e-mail services (for example, Gmail and Yahoo). Additionally, communication between individuals having different native languages has also evolved through free online translation services. All this not only involves sharing data and information, but also includes sharing of feelings.

Many theories exist on the basic human emotions [18, 19]. The widely accepted of these is the Paul Ekman's theory according to which there are six basic human emotions, namely, happy, sad, fear, disgusting, angry, and surprise. In the contexts of affective computing, the machine learning techniques are used to extract meaningful data for detecting dominant emotion. Understanding the human emotions through the text is an important research undertaking. Identification of emotions from writing pattern and text can be used for the improvement of the user interaction in a variety of contexts. Research on the detection of emotions from e-mails, tweets, short messages, and video comments (on YouTube, and so on.) is used to enhance the user experience for CQI purposes.

1.2 Scope and Applications

The scope of the present work is to recognize dominant emotion in short text samples using its in-text features coupled with the keystroke information. The short texts may consist of web links, e-mail addresses, or other attachments. However, this work focuses only on the textual data of short text and its associated keystrokes for the features extraction and emotion recognition. For this, four feature selection methods, namely, tree-based feature selection, **mutual information** (**MI**), univariate feature selection, and statistical dependency-based feature selection are utilized. For the emotion recognition, supervised learning is used with three classifiers, i.e., **Support Vector Machine** (**SVM**), discriminant analysis, and ***k*-Nearest Neighbors (*k*-NN)**. Short text is a common mode of communication for personal interaction and at times also utilized in the official correspondence. Identification of the dominant emotion in a particular communication will help in appropriate interpretation at the receiving end. This will also have other benefits from the business point of view, like, targeted advertisement, emotional therapy, and many others.

In this context, the creation of the dataset which contains the text samples and associated keystroke information is within the scope of the present work. For the creation of this dataset, a set of participants is engaged and the emotions are induced into them through displaying videos specific to each emotion. It can be seen in the past work that stimuli are chosen according to target classes of emotions as done in [20]. For the creation of IEMOCAP database, authors in [20] selected television plays in the supervision of a theater professional who then chooses plays which can convey target emotions (for example, happiness, anger, sadness, frustration, and neutral state). A similar approach has been applied in the present work to induce basic human emotions in the participants. The proposed approach is applied to both kinds of collected data, i.e., the text and keystrokes to build a single framework which can identify emotions from text and keystrokes and to compare their performance under the same conditions.

1.3 Our Contribution and Novelty

This work presents a CI-based novel framework and its associated baseline dataset for the identification of human emotion hidden in the limited text. For this purpose, initially, a dataset is collected in a real-world setting by inducing emotions in a set of human subjects in a controlled environment. Primarily, the dataset contains only text data. However, a software utility is executed while the participants write their views and the various keystroke events are also recorded, which are integrated with the final dataset. The collected database contains short length text data and keystroke information based on five basic human emotions (for example, happy, sad, or surprised) induced after watching videos. The labeling of data is based on the video contents and its comments. The labels are assigned based on a voting procedure involving five annotators. The final dataset consists of free text and keystroke dynamics of participants' affective states. Later, the CI methods are used to preprocess the dataset and extract important features that influence in identifying an emotion. The preprocessed data is utilized by the proposed framework to extract the dominant human emotion hidden in the short text. The framework includes four feature selection techniques and three classifiers for the task at hand. This proposal design six models using CI techniques to classify data into five classes and two dimensions of emotions. The proposed approach is compared with four closely related state-of-the-art methods, namely, Shikder et al. [12], Tripathi et al. [21], Alhuzali et al. [22], and Aguado et al. [51]. Two of these comparison methods are based on deep learning, Simulations show that the proposed models achieve better accuracy on the DEKT-345 × 2 dataset as compared to three other methods when applied to this dataset. There can be multiple scenarios where access to the actual data is not possible and one can only rely on the keystroke information, for example, encrypted data, password fields, and answers to secret questions, to name a few. The present contribution, i.e., the dataset and the allied methods are important from the perspective of these scenarios.

A supervised learning solution for emotion recognition from small text samples is presented here. The proposed approach identifies the dominant emotion hidden in the limited text and categorizes it into one of the five basic emotions, i.e., happy, sad, angry, surprise, and disgusting. This is achieved based on the analysis performed on the small text and its associated keystroke information. Recognizing emotions from small text is a challenging task and to the best of our knowledge, a dataset that contains labelled instances with all of the abovementioned emotions is not publicly available. To address this issue, the current work creates such dataset for the utilization by the pattern learning module. Based on the abovementioned details, the key contributions of this work are as follows:

- Presentation of a supervised learning-based framework to identify dominant emotion hidden in small text.
- Utilization of in-text features for the prediction of emotion.
- Extraction and utilization of keystroke information for the prediction of emotion.
- A novel balanced and labeled dataset containing five basic human emotions.
- Contribution of a non-posed dataset covering human emotions.
- Contribution of a novel balanced and labeled dataset containing five basic human emotions using text and keystrokes information to be utilized by the AI community to train existing and build new models for the emotionally intelligent computing systems.

The key novelty of this work is the utilization of both in-text textual features and the keystrokes information to identify the dominant emotion hidden in the limited text. A few of the past contributions based on deep learning methods, like [23] and [21], consume complete data samples at once as a batch and predict the emotion. A drawback of such methods is that they do not let the user specifically know what data item has been used for the prediction purpose and the attributes that play a vital role in the prediction task remains unknown. These methods have their own advantage, however, they remain a black box and there is nothing much for the user to manipulate. Another novelty aspect of the present work is that it provides the user with the flexibility to select and extract a variety of features from the dataset and experiment with them to see which ones perform better. However, for the sake of a fairer assessment, the proposed approach is compared with two of the deep-learning methods for the same task. Yet another important innovative aspect of the present work is the contribution of a labeled dataset consisting of short text and the allied keystrokes information for the research community. Another novelty of the present work is that the dataset contributed by this work is of non-posed nature.

The rest of the article is organized as follows. Section 2 presents the literature review. Section 3 contains a detail of the experimental setup and method for data acquisition. Section 4 presents the proposed approach. Section 5 lists the results and discussion. Finally, Section 6 concludes this work.

2 LITERATURE REVIEW

Human emotions identification is an important research task with many applications. The accuracy of the underlying emotion identification modules is dependent on the training they receive through the collected/available data. Most of the effective databases in previous works are designed in guided and acting scenarios. A few of them are available online for further research work. Emotion detection systems in most of the previous works are based on sensors and other wearable peripheral devices, which are costly and difficult to use in the data collection procedure for the participants [15, 20, 24]. This section lists the previous approaches for this task.

Gupta et al. [25] detect emotion from e-mails which are sent by customers on a specific product to the correspondence center. They provide valuable feedback to improve the contact center process and to enhance customer retention. Their work describes a method to extract salient features and to identify emotional e-mails. The salient features include frustration, dissatisfaction, and threats to either leave the business or to take any legal action. Seo et al. [26] present an automatic method to identify the emotions of music listeners. They classify music according to the emotional range of listeners. Classification is performed using multiple regression methods and comparative analysis is performed using other classification algorithms such as **random forest** (**RF**), deep neural network, k-nearest neighbor, and SVM. Wolff et al. [27] investigate the detection of emotions in **Borderline Personality Disorder** (**BPD**) patients compared with healthy individuals. In their study, 30 female BPD patients and 28 healthy female participants are asked to enter data on an hourly basis for a 24 hours duration. The results indicate that the emotion identification of BPD patients is challenging as compared to healthy persons. Twitter generates Big Data on different topics. Many of the tweets contain opinions on various ongoing sports. The work in [28] presents a convolutional neural network architecture for emotion identification on Twitter messages during sports events of the 2014 **Fédération Internationale de Football Association** (**FIFA**) world cup. Their network use pre-trained word embedding on large text corpora. The work in [29] develops an emotion recognition method of employees for the improvement of the organization's processes. They record front face images of the employees as they enter for work in the organization.

Keyboard and mouse movements are sensor-free and hardware-based sources for affective states detection. Different features of keys are used for the analysis of the human affective state. Shikder et al. [12] propose a method for emotion detection and user identification using keystrokes and mouse movements. They induce emotions in participants by showing them nine video clips of three emotion dimensions. Features extracted from mouse click are: average mouse left click, average mouse right-click, average mouse double click, average mouse scroll, average cursor x-distance, and average cursor y-distance. Features recorded from keystrokes include: average key down to uptime, average key up to down time, average key down to down time, average regular key press, average enter key press, average function key press, and a few others. The authors use bounded k-mean clustering and k-NN classifier. Salmeron-Majadas et al. [13] propose a method to analyze behavior of users while using the keyboard and mouse in real-life applications. Their goal is to evaluate the approach in the education system; therefore, they collect data from essays written by the participants. To capture the data, they use the MOKEETO tool, which is a mouse and keyboard-based essay writing software. It is also a keylogger and a mouse tracker. They collect data in two sessions. The first session is held in a high school where 27 participants participated and the second is in **Universidad Nacional de Educación a Distancia** (**UNED**) with 14 participants. They extract features related to the keyboard's keys, key independent features of the skeyboard, features on mouse movements, and features related to the given task. Classification is performed based on arousal and valence attributes. Binary classification, i.e., positive and negative emotion detection is performed in their work. While adding a third class, i.e., neutral emotion, 80.6% (maximum) accuracy is reported [13]. The work in [30] asks the participants to perform three programming tasks of varying difficulties to analyze their three affective states, i.e., positive, negative, and neutral. An accuracy of 52.9% is achieved for three classes and 74.1% for two classes using a feed-forward neural network. The frustration of programmers while programming is analyzed in [31] using keystrokes dynamics and mouse clicks. An accuracy of 73% is achieved using a Bayesian network and 55% utilizing naïve Bayes classifier. The work in [32] presents a machine learning-based solution to identify the dominant emotion hidden in the e-mail text. Their study considers six basic emotions and a total of 15 text-based features are extracted from the e-mail. The accuracy reported in their work is 83% for the emotion identification using the ensemble of multiple classifiers. The authors used a custom build dataset for their experiments and utilized clustering techniques to verify the number of emotions available in the complete data before continuing with the classification task. Khare et al. [33] present a convolutional neural network-based approach to automatically retrieve features for the classification of emotions. Their proposal utilizes the EEG signals for this purpose. Performance evaluation of their work is done through accuracy, precision, Mathew's correlation coefficient, F1-score, and false-positive rate. The proposal in [34] creates a computational model to identify the conceptual meaning of the words utilized to express emotions. For this, they present the meaning of emotion words as interval **type-2 fuzzy sets** (**IT2 FSs**). Akhtar et al. [35] present a stacked ensemble method for recognizing the degree of emotion intensity. This is done by joining the outputs obtained from several deep learning and classical feature-based models using a multi-layer perceptron network. Their model is evaluated for emotion recognition in the generic domain and for sentiment analysis in the financial field.

It is evident from the above-mentioned literature review and to the best of our knowledge, most of the available affective databases are based on sensory data [15, 20] or audio/visual cues [16, 36, 37]. Whereas, the majority of the past work [38–44] is on acted data [37] which cannot be conveniently used in many real-life scenarios. The focus here is to develop a non-acted non-sensor database and present an allied framework to identify the dominant emotion hidden in short texts. For this, text-keystroke sources are used for affective states detection.

3 SETUP AND METHOD FOR DATA ACQUISITION

The DEKT-345 × 2 dataset has been collected in the research lab's premises. An experimental setup is designed to achieve the research goals, including: (a) induction of emotions by watching short video clips, (b) providing an environment to the participants that is relatable to real-world scenario, (c) minimum possible engagement of the participants with hardware tools, and (d) be able to analyze subjects' emotions after experiencing a real-life video. To analyze emotions induced in the participants after watching a scene related to real-life events, the participants were asked to express their feelings in a maximum of 150 words using English as a language. The dynamics of keystrokes during typing are recorded for experiments. The goal is to collect an equal number of samples per class from each participant. This enables to make a comparison for emotion detection between subjects, text, and keystrokes. This work utilizes videos to induce emotion in the participants before asking them to give samples. Initially, the participants were received in the room (that is, the research group's general-purpose computing laboratory with audio and video facilities). As a first step, all participants were briefed about the complete data collection process. The next step was the collection of demographic information from the participants. The participants inducted in the data collection process came from different backgrounds and may already be in a specific emotions (for example, happy, sad, or angry). This would simply add noise to the recorded samples. To mitigate this factor, the participants were brought to a baseline before inducing the emotion through videos. For this, the participants were requested to walk towards the wall of the room and stand there (while facing the wall) for one minute. This helped to bring the participants to the baseline. Afterward, they were requested to be seated and the emotion was induced through videos. As the video was completed, text samples were taken from the participants. During this, the software utility recorded keystroke information. Once the samples were taken, the participants were also asked to self-report their current emotion. The self-reported emotion was available to the annotators while assigning the final labels to the samples based on the recorded text. This work do not explicitly use the self-reported emotion as a label because the primary objective here is to identify emotion based on the text (and its keystroke information). There is a possibility that a participant may not provide the appropriate amount (or quality) of the sample (i.e., the written text) and only gives the self-reported emotion. Therefore, the annotator utilizes the self-reported emotion and the sample to assign an appropriate label to each record. For final labelling, this work utilized five annotators. Each annotator went through the user-provided content plus their self-reported emotion and assigned a label (i.e., one of the emotions). Final labeling was decided based on the majority vote by the five annotators. Figure 1 shows the overall data collection procedure.

The keylogger program used here for recording keystrokes is written in the Python programming language to trace the time difference between key press and key release events. Participants executed this program before the data collection phase. As the software starts, it asks the user to

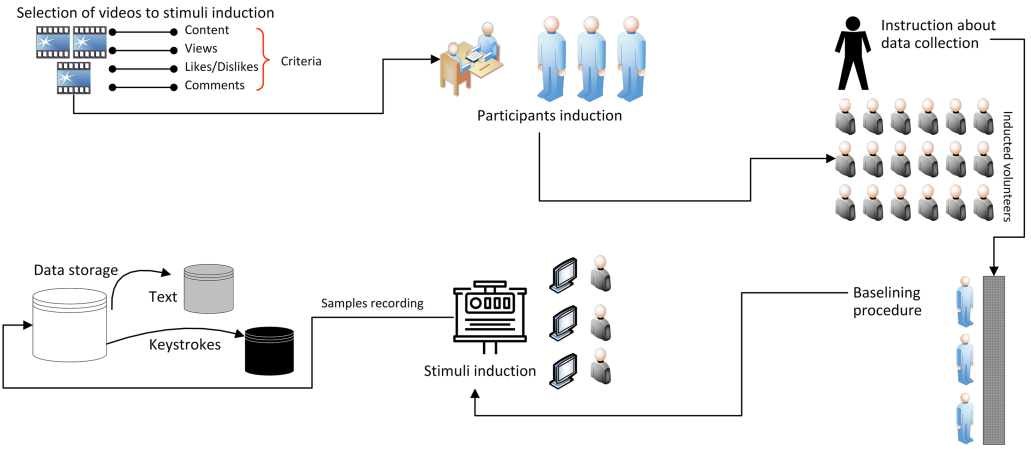

Fig. 1. Data collection procedure overview.

Table 1. Content and Target Emotion of Stimuli

Content	Views (as of July 2021)	Label	URL
Victory of Pakistan World Cup 1992	93,302	Happy	https://www.youtube.com/watch?v=rL4xKdqyMgY
Tourism in Pakistan	482,400	Happy	https://www.youtube.com/watch?v=nXwijt1EUDo
Illegal Immigration	1,666,572	Disgusting, Fear	https://www.youtube.com/watch?v=oyBRr8h4ZrI
Destruction of natural resources	51,428,995	Disgusting	https://www.youtube.com/watch?v=WfGMYdalClU
Pathetic education system and mental pressure	27,447,710	Fear, Sad	https://www.youtube.com/watch?v=BE4oz2u6OHY
Syria war's causes	5,080,897	Anger	https://www.youtube.com/watch?v=K5H5w3_QTG0
Syria's affected people's condition	1,093,123	Fear, Sad and Anger	https://www.youtube.com/watch?v=jdKHVnHTXkU
Mystery of Bermuda triangle	28,783,649	Surprise	https://www.youtube.com/watch?v=q_5n7URd2Gk
Passion of a scientist to observe lava lake	12,503,345	Surprise	https://www.youtube.com/watch?v=egEGaBXG3Kg

enter the participant name, in the result of this, an excel file with the name of that user is generated in a separate folder. This enables to keep track of each participant's data. Five videos related to five different emotions (i.e., happy, sad, angry, surprise, and disgusting) are shown to the participants for emotion induction. These videos are chosen on the basis of their content, number of views, number of likes, number of dislikes, and comments. The videos are shown to the participants using a multimedia projector on the front wall of the research lab. Table 1 lists details about the videos.

To achieve natural emotions and without the pressure of time or other such constraints (e.g., the use of specific language or dictionary), the participants are free to write a text about what was in their minds. Past literature mentions various approaches for this, for example, in [39], the authors use fixed data to write about after selecting the emotion class. Similarly, in [42], the analysis is based on time pressure and without any time limit. However, the focus here is on the real-life scenario and to identify natural emotions that arise in one's mind. A few samples of the short text captured against each emotion are listed the following.

Happy: "The video is awesome. The woman has portrayed the actual beauty of Japan. This is something that is very wrongfully depicted. Japan and its people welcome all such tourists to come and explore this holy land and cherish this wonderland."

Fear: "Do not follow others like sheep. Do not categorize all children in one category. Every child has his own passion. Nobody is perfect. Nobody burns bad. It is us who made them."

Table 2. Databases Consisting of Human Emotion Information

Database	Language	Type	Annotation	Type of Data	Emotion Stimuli	Number of Stimuli, (and Subjects)
BAUM-1 [16]	Turkish	Un-Posed	SBE, 2 NBE, 3 MS	Audio, Visual	Videos Clips	273 (acted), 1,184 (spontaneous), (31)
DEAP [14]	–	Un-Posed	5 NBE (A,V, Like/Dislike, Dominance, and Familiarity)	EEG, Peripheral Physiological Signals (GSR, respiration amplitude, skin temperature, ECG and so on.)	Video Clips	40, (32)
IEMOCAP [45]	English	Posed	Anger, happiness, sadness, neutrality/Valence, activation, and dominance	Videos, Speech, Head Movement and Head Angle Information, Dialog Transcriptions, Word level, Syllable level and Phoneme level alignment	Scripts for actors	-, (10)
eNTERFACE [36]	English	Posed	SBE	Audio and Visual	Video Clips	1166, (44)
Emo-DB [37]	German	Acted	happy, angry, anxious, fearful, bored, and disgusted	Speech	10 German utterances	10, (10)
DEKT-345 × 2 (current)	English	Natural	5BE	Text and Keystrokes	Videos	5, (69)

Anger: "Children suffer the most due to war, they are robbed of their futures and dreams. It is honorable that some doctors put their lives at risk to protect them, but it is a shame that hospitals are bombed. Although we cannot physically be there to help them, we can still donate and raise awareness about the issue."

Disgusting: "The video depicts the evolution of human life and how humans have used animals and various natural resources to give their life shape and sense of direction. They kept on utilizing the resources until the world turned into huge chaos and piles of rubbish. The future of this world has been demonstrated in this short video."

Surprise: "Person having an adventurous mind like me would be very curious to know what would happen next and as the lava was bubbling and trying to splash out was in shock many times that it's going to hit the man. What would happen next? Is he going to survive? And when he got the got back safely at that instant right after he was back lava splashed and I uttered the words "Thank God" he backed out at the right and exact moment."

There have been multiple affective datasets presented in the past. A summary of these databases is given in Table 2. BAUM-1 is an audio-visual database in the Turkish language. It is an un-posed dataset with five basic emotions, two non-basic emotions, including boredom and contempt, and three mental states, i.e., unsure, concentrating, thinking, and bothered. Another database is created in [14] for emotion analysis using physiological signals, called a **Database for Emotion Analysis using Physiological** (**DEAP**) signals. It is an un-posed dataset consisting of five non-basic emotions. Video clips are used as emotional stimuli in [14] while 32 subjects participated in data collection. IEMOCAP dataset presented in [45] is a posed dataset in the English language. A total of 10 individuals participated to act for this dataset. Written scripts are used as emotional stimuli. The eNTERFACE dataset is created in [36]. It is a posed dataset in which 44 subjects posed in audio and video media after watching video clips. This dataset consists of data of seven basic emotions. A dataset named Emo-DB is created in the German language which consists of speech data of 10 participants. It consists of data of happy, angry, anxious, fearful, bored, and disgusted classes of emotions. As evident from the past affective databases mentioned here, the dataset presented in this work is a novel affective database in a sense that it is an un-posed dataset consisting of data of text and keystroke dynamics. The past contributions are either limited to text or posed contents. Whereas, here, the non-posed aspect is the focus and collection of keystrokes information of the associated text data is the primary novelty.

There is always a tradeoff between the quality of data and its quantity. The dataset in this work is collected without the use of any written transcripts or through actors. This enables to collect the natural writing patterns of the subjects instead of relying on some guided response. Such type of data always remains limited in quantity, but the data quality remains better. The same trend is also observed in the closely related past literature, like [14, 16, 36, 37, 45] where they collect data using stimuli from subjects in the form of written scripts, EEG, ECG, body parts' movements that could result in a larger number of samples. Whereas, the text and keystrokes data collected in [39], is in the form of comments which is limited in size, i.e., a response of 25 volunteers is collected only. The solution proposed in this work is tested specifically for the limited text scenario. Apparently small data may lack the transferring ability, however, the recent developments in deep learning demonstrate the reasonable performance of model like **Log Expected Empirical Prediction** (**LEEP**) [52, 53] even on small or imbalanced data settings.

The present work considers the emotions identified in the Paul Ekman's theory. He identified six basic emotions according to the isolated culture of people from the Fori tribe in Papua New Guinea in the year 1972. He also ordered them as anger, disgust, fear, happiness, sadness, and surprise. Therefore, the present work has adopted the same order for the induction of considered emotions into the participants to collect the data. A few past contributions, like [46] has used eight emotions, namely, happy, calm, love, positive surprise, negative surprise, angry, sad, and afraid. The two additional emotions considered in [46] are calm and in love. Since these two are not mentioned in the Paul Ekman's theory as basic emotions and inducing these two is also challenging, therefore the present work continues with the five emotions based on the majority of the past works and utilizes the standard theory on human emotion [47–50].

3.1 Participants

The data is collected in two sessions. The first session was conducted during summer vacations. Where 48 graduate and undergraduate students (75% males and 25% females) were asked to participate in the data collection phase. The second session was held in spring, where 21 undergraduate students (80% males and 20% females) participated in the data collection process. As mentioned earlier, the participants are shown videos related to each of the emotions. This serves as stimuli to induce an emotion into the volunteers. This work has opted for one video per emotion so that each volunteer watches a single video against each emotion. This is done to remain aligned with the strategies mentioned in the past literature. One could opt for displaying more than one videos for a particular emotion, for example sad, however, the contents, storyline, and presentation of multiple videos will not be the same and shall result in contradictory or inaccurate emotions being induced due to varied contents.

3.2 Hardware

The hardware used in the data collection phase consisted of desktop computers with Intel core i7 processor, 2 GB RAM, 64-bit operating system, USB keyboard of BELKIN, a USB mouse of A4TECH brand with two buttons and a scroll, and a ViewSonic monitor.

3.3 Software

A keylogger is developed in Python programming language to trace keys and time differences between their press/release events by the participants. The participants executed the keylogger in Anaconda/Python shell which traced various keystroke events. After running the program, an excel file is generated with the identity of the participant and her collected data.

3.4 Labeling

The availability of ground truth plays an important role in any learning task, especially those which involve supervised learning activity. The same is true in the present case. As mentioned earlier, this work utilizes the video display mechanism for the induction of emotion into the participants. A video is displayed against each of the emotion. Initially, the videos are shortlisted based on their title and contents to be reflective of the particular emotion under consideration and also based on the number of views. Afterward, the labeling of shortlisted videos is done based on the content of videos and voting by a set of annotators (odd in number). The method utilized here is motivated by the labeling procedure mentioned in [16]. This enables to have the emotion induction video reflective of the particular emotion under consideration and collect reliable data. The final set of labeled videos are then shown to the participants to express their views about the video. During the labeling procedure, each annotator could watch the clip as many times as she desired. She selected the emotion that is dominant in the clip as the final label. After collecting the labeled data from all the annotators, the final labeling is decided based on the majority vote.

4 PROPOSED APPROACH

This proposal utilizes CI methods to extract the dominant emotion from short texts. For this, as part of the complete framework, a detailed dataset (DEKT-345 × 2) is collected. The usefulness and nature of data are evaluated by conducting effective multimodal experiments on the DEKT-345 × 2 dataset. The motivation here is to develop a system that can identify the emotional state of a person who is writing the short text. The system is designed in such a way that participants do not need to wear any peripheral devices. They are asked to write free text which comes to their mind after being induced with the stimulus. The focus in this work is on the analysis of keystrokes dynamics and text data to detect basic human emotions as described by the Ekman's theory.

This work presents a methodology on (a) emotion induction procedure, (b) data collection environment, (c) model generation to analyze the effectiveness of keyboard keys in effective computing, (d) model generation using machine learning techniques for emotion detection using keystrokes, (e) model for emotion detection using text data, and (f) emotion detection using a hybrid approach. Most of the previous work is based on keystrokes and mouse movement, as in [12, 13, 25]. Machine learning techniques are applied in the literature for emotion detection using keystrokes [24, 42, 43]. The primary focus here is on keystrokes and text analysis to evaluate real-life scenarios and the use of machine learning techniques in the experiments.

4.1 Preprocessing and Feature Extraction of Text and Keystrokes Features

The feature selection is applied to the complete dataset instead of restricting it to the training or test sets utilized for the classification purpose. This includes both the text data and its keystroke information. Once the final set of features is selected, these are extracted from the data and the training of the classifiers is performed. For training, 80% of data is used and testing is performed on 20% of data. Training and validation are done for 20 different combinations of training and validation data while testing is performed on unseen data for which class labels are stored separately to compute confusion matrix.

Preprocessing: As mentioned earlier, the data is collected in the research group's lab in the raw form. In the case of keystrokes, some of the values were missing from the data. To fill-in the missing values, the nearest value approach is used. In this approach, the missing value is replaced by the nearest non-missing value.

Feature extraction: Later, time and frequency-based features are extracted from the keystrokes data. Typing the speed of participants is a subjective matter. Some persons may be used to thinking about the content of the text for a long time while others may finish writing the text quickly. To consider this aspect of persons, we extracted the feature of typing speed and the difference between the press and release time of a key. Frequency-based features are used to represent a person's behavior which is affected by sentiment changes or stress level. A list of these features is

Table 3. Features Extracted from Keystrokes

Feature No.	Feature
1–26	(T_t) Total time duration of each letter (a to z)
27–52	(T_a) Average time duration of each letter (a to z)
53–78	(σ) Standard deviation of time duration of each letter (a to z)
79	(S)Typing speed
80	(F_s) Frequency of space key
81	(F_r) Frequency of right shift key
82	(F_l) Frequency of left shift key
83	(F_{rc}) Frequency of right control key
84	(F_{lc}) Frequency of left control key
85	(F_{err}) Frequency of error
86	(F_e) Frequency of enter key
87	(F_u) Frequency of up control key
88	(F_d) Frequency of down control key
89	(T) Total time taken by user

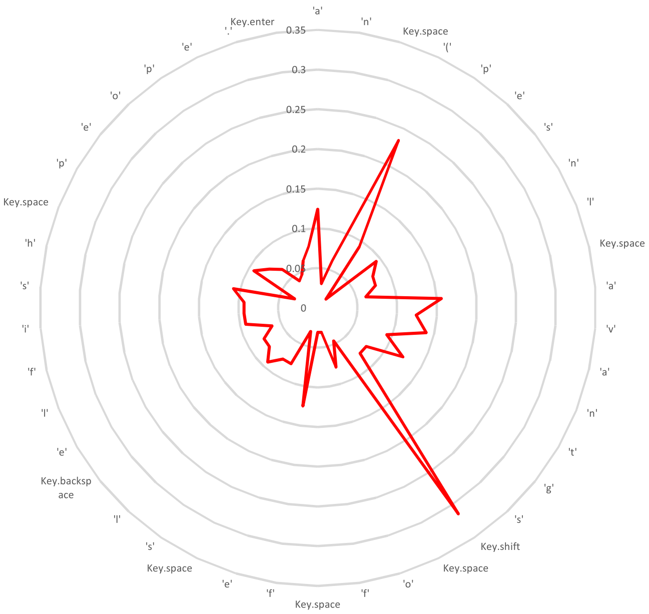

Fig. 2. Example of keystrokes of a participant's feelings on a video.

shown in Table 3. Features from 1 to 26 are measured by calculating the sum of time in seconds for which a letter was pressed in a full review of a video. Consider the following example.

Example 1. "Wars are based mostly on greed and (personal advantages) of few selfish people".

Figure 2 display the keystrokes related information collected against 84 keys for a single sample in this work. For a different sample, the number of keys against which the data is collected may vary depending upon the actual text written by a person. For example, in case of the data visualized in Figure 2, the keys enter, shift, "i", "d", and "e" are among the 84 keys whose data is collected. The figure shows the time duration in seconds for which the particular key is pressed. Like, the key "a" is pressed for 0.171s and "r" is pressed for 0.124 s. According to Figure 2, in *example-1*, time of the letter "a" will be, 0.171 + 0.109 + 0.124 + 0.14 = 0.544. Features from 27 to 52 are calculated by taking the average of time duration of a letter, like in *example-1*, an average time of "a" = 0.544/4 = 0.136. Features from 53 to 78 are calculated by taking the standard deviation of time duration (σ) of a letter, as

$$\sigma = \sqrt{\sum (x - \overline{x})^2}\,/n \tag{1}$$

Feature number 79 is regarding typing speed of a user. It is computed by dividing the length of letters pressed and a total time of the letters in an expression's keystrokes using (2). Typing speed (T_t) considering *example-1* will be 7.3196 letters/second, i.e.,

$$Tt = \frac{length\,(data)}{total\ time}. \tag{2}$$

Frequency of space key usage (F_s) is computed using (3), in *example-1*, it will be 0.1566, i.e.,

$$Fs \ = \ \frac{No.of\ space\ keys}{total\ letters}. \tag{3}$$

Frequency of right shift key usage (F_r) is calculated using (4), for *example-1* it is 0, i.e.,

$$Fr \ = \ \frac{No.of\ right\ shift\ keys}{total\ letters}. \tag{4}$$

Frequency of left shift key usage (F_l) is calculated using (5), considering *example-1*, it will be 0, i.e.,

$$Fl \ = \ \frac{No.of\ left\ keys}{total\ letters}. \tag{5}$$

Frequency of right control key usage (F_{rc}) is computed using (6). In case of the sample in *example-1*, it will be 0, i.e.,

$$Frc \ = \ \frac{No.of\ right\ control\ keys}{total\ letters}. \tag{6}$$

Frequency of left control key (F_{lc}) is computed using (7), in *example-1*, it will be 0.0241, i.e.,

$$Flc \ = \ \frac{No.of\ left\ control\ keys}{total\ letters}. \tag{7}$$

The frequency of error (F_{err}) is computed using (8), for the data in *example-1*, it will be 0.0241, i.e.,

$$Ferr \ = \ \frac{No.of\ backspace\ keys}{total\ letters}. \tag{8}$$

The frequency of enter key (F_e) is calculated using (9), considering *example-1*, it will be 0.0361, i.e.,

$$Fe \ = \ \frac{No.of\ enter\ keys}{total\ letters}. \tag{9}$$

Frequency of up control key (F_u) is calculated using (10), in example-1, it will be 0, i.e.,

$$Fu \ = \ \frac{No.of\ up\ control\ keys}{total\ letters}. \tag{10}$$

Frequency of down control key (F_d) is computed using (11). For the given example, it will be 0, i.e.,

$$\mathrm{F}_d \ = \ \frac{No.\ of\ down\ control\ keys}{total\ letters}. \tag{11}$$

The total time (T) taken by user is calculated using (12). In *example-1*, it will be 11.47, i.e.,

$$T = sum\ of\ time\ consimed\ to\ type\ all\ letters. \tag{12}$$

The values of some of the features are in different ranges. To normalize all features in a specific range, these are scaled between 0 and 1 using

$$a = x - \min(x), \tag{13}$$

$$b = a./\max(x). \tag{14}$$

where x is the feature vector.

ALGORITHM-1: *Tree-Based Feature Learning Procedure.*

Procedure: Random Forest

```
Input: Training set ($S$), Features ($F$), No. of trees ($T$)
H←Null
for each $i$ from 1 to $T$
        $S_i$ ← A bootstrap sample from $S$
        $h_i$ ← $Tree\_learner$ ($S_{i,}$ $F$)
        $H \leftarrow H\ U\ h_i$
end for
return $H$

Tree_learner (S, F)
For each node:
        f ← small subset of F
        Split based on best feature in f
return $tree_{learned}$
```

Text-based features are extracted using scikit-learn software available in the Python programming language. The **Term Frequency and Inverse Document Frequency** (**TFIDF**)-based features are extracted using (15). It assigns a numerical value to a word, which shows its importance in a corpus document. The term frequency is the count of the term that appears in a document and inverse document frequency is the measure of information a word provides. It is obtained by taking a *log* of the total number of documents with the number of documents containing the term.

$$IDF\,(t,d) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right), \tag{15}$$

where N is the number of documents, t is the term in document d; D shows all documents in a corpus. Unigrams and bigrams are extracted from the text data. There are a total of 9,905 unigrams and bigrams in the captured dataset. To address this concern of a very large feature set, the overall feature dimensions are reduced using multiple feature selection techniques to avoid overfitting problems in the classifiers' training phase. Thus dimensions of features get reduced and diverse after applying feature reduction/selection techniques.

4.2 Feature Selection

Four feature selection techniques are used to extract the important features required for the experiments. These include tree-based feature selection, MI, univariate feature selection, and statistical dependency-based feature selection.

Tree-Based Feature Selection: The tree-based estimator is used to compute feature importance and discard irrelevant features. Extra trees are used to compute the importance of the features. Extra

Table 4. Top 10 Features Selected from Text and Keystrokes Data Using MI

Text data		Keystrokes data	
Feature	Weight	Feature	Weight
Pakistan	0.357	Average time of a	1.0123
Bermuda	0.2375	Mean of q	0.9591
Syria	0.2344	Mean of x	0.5582
Triangle	0.2288	Mean of j	0.5492
Bermuda triangle	0.2115	Standard deviation q	0.2501
Children	0.1682	Standard deviation z	0.2358
War	0.1552	Mean time of k	0.1931
In Syria	0.1517	Standard deviation of r	0.178
In	0.1389	Standard deviation of k	0.1654
Of	0.1369	Total time of f	0.1638

tree classifiers are used for extremely randomized trees. It builds multiple trees and divides its nodes into random subsets of features. Feature selection using RF is a kind of embedded method which combines the property of both filter and wrapper approaches. The same is adopted here as a tree-based feature selection technique. The RF is based on bagging and random feature selection. Bagging is a resampling procedure that creates bootstrap instances by randomly sampling with replacement from the original training sample. RF is essentially an ensemble of trees where each grows based on a different bootstrap sample. A RF algorithm consists of multiple **Decision Trees** (**DT**). The bunch of DTs classify the forest generated by the FR procedure which is trained through bagging/bootstrap aggregating procedure. The performance of randomization can be improved by adjusting the parameters. Results show that the tree based-method gives important features that yield better accuracy as compared to others.

Mutual Information: MI measures the mutual dependence of two variables. It calculates the information gain of one random variable by observing the other random variable. The MI of feature values with respect to class labels is evaluated using (16). Table 4 shows the features selected using MI from the DEKT-345 × 2 dataset.

$$MI = \sum_{x \in X} \sum_{y \in Y} p\left(x, y\right) \log \left(\frac{p\left(x, y\right)}{p\left(x\right) p\left(y\right)} \right). \tag{16}$$

where x and y are random variables, x represents a feature vector and y is the vector of class labels. $p(x, y)$ is the joint probability of x and y. The values of $p(x),\ p(y)$ are the marginal probabilities of x and y, respectively.

Univariate Feature Selection: Univariate selection uses the chi-squared ($\chi 2$) statistical test to select k best features. Chi-squared test is used to determine whether there is a significant difference between the actual frequencies and observed frequencies in one or more categories. Results are shown in Table 5.

Statistical Dependency-Based Feature Selection: The statistical dependency method measures the dependence of values of feature with its label. The statistical dependence between the quantized feature values x and the class labels y are measured using (17). Results are shown in Table 6.

$$SD = \sum_{x \in X} \sum_{y \in Y} p\left(x, y\right) \frac{p\left(x, y\right)}{p\left(x\right) p\left(y\right)}. \tag{17}$$

Larger value of SD indicates a higher dependency between feature values and the class labels.

Table 5. Top 20 Features Selected Using Univariate Feature Selection

Text data		Keystrokes data	
Feature	Score	Feature	Score
Pakistan	3.6	Total time	0.27
Bermuda	2.58	Sum of e	0.08
Triangle	2.24	Sum of r	0.069
Bermuda triangle	2.17	Sum of o	0.066
Dangerous	1.4	Sum of s	0.052
Team	1.34	Sum of w	0.041
She	1.29	Sum of t	0.036
Mystery	1.25	Sum of a	0.035
Gilgit	1.22	Typing speed	0.035
Of Pakistan	1.16	Sum of n	0.0335
World cup	1.06	Sum of m	0.033
Cup	1.05	Sum of h	0.028
Confusing	1.04	Sum of p	0.028
Confusing video	1.04	Sum of c	0.026
Very confusing	1.04	Sum of k	0.025
Trip	1.04	Sum of i	0.022
Beautiful	1.02	Sum of f	0.021
Human	0.97	Sum of y	0.021
Beauty	0.96	Sum of u	0.018
Beauty of	0.92	Sum of z	0.017

The dimension of text features is larger than that of the keystroke attributes. Apparently, this may reduce the influence of keystroke features. To address this aspect, the feature dimensions of text data's attributes in this work are reduced using feature selection techniques to avoid overfitting problems and to combine them with the statistical features pertaining keystrokes.

4.3 Classification and Fusion of Text and Keystrokes Models

Three classifiers, namely, SVM, discriminant analysis, and *k*-Nearest Neighbors (*k*-NN) are used here for the classification of text and keystrokes features. The SVM classifier uses a linear kernel as it performs well in case of the high dimensionality of the features. Text and keystrokes are classified on a one-against-one multiclass classification approach. The discriminant analysis classifier has multiple types. However, this work uses either linear or diaglinear category of the discriminant classifier due to its better performance on small data. A feature level fusion technique is applied to evaluate the effectiveness of text and keystrokes for the detection of emotions. The feature vectors of text and keystrokes modalities are denoted as x_1 and x_2, respectively. These are fused, as shown in

$$y = x_1 + x_2. \tag{18}$$

Feature selection techniques mentioned in the previous section are applied on y feature vectors, and classification is performed on the resultant selected features. Classification is performed on text model, keystrokes model, and hybrid of the text and keystrokes models. It has been done for all five basic emotions and for the pleasure dimension of emotions. Here, the pleasure dimension is selected because the collected data lies in this dimension. Robert Plutchick categorized happy class in pleasure dimension and angry, disgusting, and fear classes in the displeasure dimension.

Table 6. Top 20 Features Selected Using SD

Text data		Keystrokes data	
Feature	Score	Feature	Score
Pakistan	0.5695	Mean time of z	1.41
Bermuda	0.3952	Mean time of q	1.23
Triangle	0.3812	Mean time of x	0.72
Bermuda triangle	0.3536	Mean time of j	0.68
Syria	0.3139	Standard deviation of time of q	0.31
Children	0.2367	Standard deviation of time of z	0.29
In Syria	0.2049	Mean time of k	0.24
War	0.1981	Standard deviation of time of r	0.22
In	0.1861	Standard deviation of time of c	0.22
Of	0.1824	Standard deviation of time of k	0.217
Of Pakistan	0.1821	Total time of o	0.216
Team	0.1769	Standard deviation of time of x	0.21
The	0.1724	Total time of m	0.2044
Dangerous	0.1663	Mean time of b	0.2041
Their	0.1654	Total time of f	0.2014
People	0.165	Standard deviation of time of f	0.2011
To	0.1639	Total time of p	0.1934
Mystery	0.1569	Total time of h	0.1895
Cup	0.1538	Frequency of space Key	0.1864
Worldcup	0.1538	Standard deviation of time of l	0.1849

Whereas, the surprise class is also categorized in the negative or displeasure domain. This categorization is shown in Figure 3.

5 BASELINE AFFECTIVE TEXT-KEYSTROKES EXPERIMENTAL RESULTS

Experiments are conducted on DEKT-345 × 2 dataset, which consists of 345 text samples and 345 keystrokes samples of the five basic human emotions (total of 690 samples). Performance on keystrokes dynamics is quite challenging as it contains individual character rather than a phrase and their time duration for which the keys have been pressed. Identification of significant features of keystrokes and text is also important for an efficient system development. The five basic emotions here are also later categorized into two dimensions of pleasure and displeasure. Experiments using four feature selection techniques and three classifiers are performed on the DEKT-345 × 2 dataset. Their results are summarized in the following.

5.1 Results on DEKT-345 × 2 Dataset for Five Class Problem

Three classifiers, namely, SVM, discriminant analysis and *k*-NN are used with each of four feature selection techniques (i.e., tree-based, MI, univariate feature selection, and statistical dependency-based feature selection). The linear kernel is used with SVM as it performs better in comparison to the other kernels. Linear discriminant type is applied in discriminant analysis because the feature dimensionality is high here. Therefore, the linear kernel and linear discriminant types are good to classify such data. The results of the models with the best accuracy are shown in Figure 4. It shows the accuracies achieved on keystrokes data for five affective states. It can be seen that SVM attains the best accuracy of 65.22% when features are selected using the tree-based technique. The discriminant analysis yields 57.97% accuracy using the MI technique of feature selection,

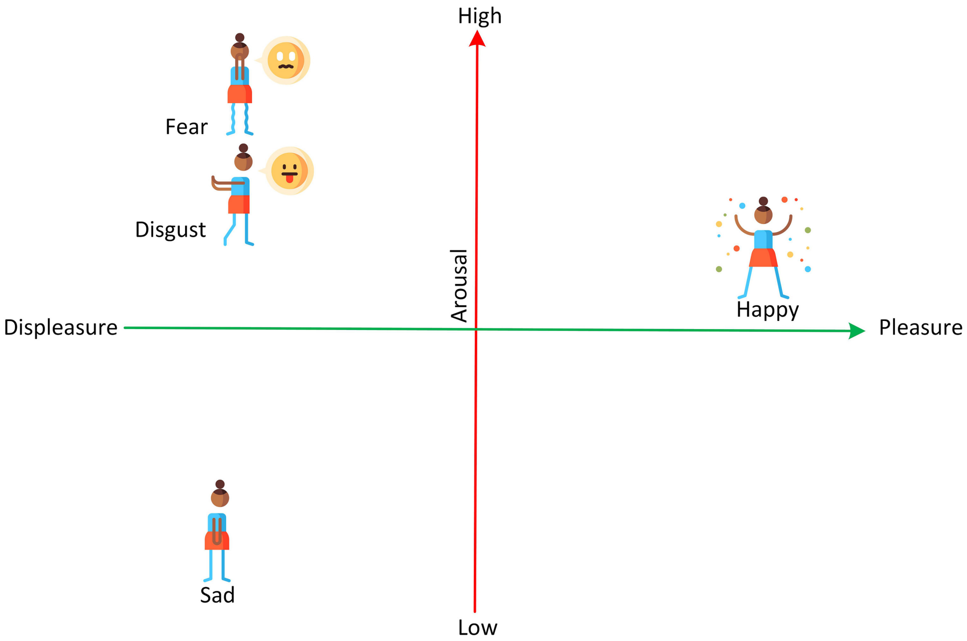

Fig. 3. Pleasure and displeasure dimensions of emotions (happy, fear, disgust, and sad).

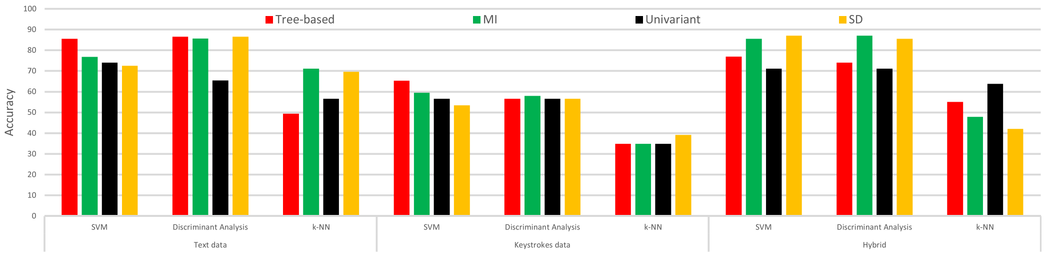

Fig. 4. Accuracy achieved on DEKT-345 × 2 database, five affective states.

k-NN has the lowest accuracy as compared to the other two classifiers. It achieves 39.13% accuracy using features selected after the statistical dependency-based feature selection method. Results also show that the class *angry* has the highest recognition rate (92.31%), while the *surprise* class has the lowest recognition rate (28.57%) fors keystrokes data of the five emotions.

Accuracies achieved on text data extracted from DEKT-345 × 2 database for five states of emotions are also given in Figure 4. It can be seen from the results that 86.53% accuracy is achieved for this model using discriminant analysis. Both tree-based and statistical dependency-based feature selection techniques perform well on text data. The SVM classifier achieved 85.51% maximum accuracy with tree-based technique and *k*-NN yielded 71.04% maximum accuracy on the features selected by using MI technique. Analyzing the per-class accuracy reveals that the recognition rate of *surprise* class is highest, i.e., 100%. Accuracy of class *happy* is 91.67%, for *angry* class it is 90%, *disgusting* and *fear* classes have 75% and 71.42% accuracies, respectively.

This work proposed a hybrid approach for emotion detection from text and keystrokes data. As shown in Figure 4, the hybrid model yields 86.95% accuracy for five classes of emotions which is greater than both text and keystrokes models. Feature selection techniques are applied after the fusion of text and keystroke features. SVM with statistical dependency-based feature

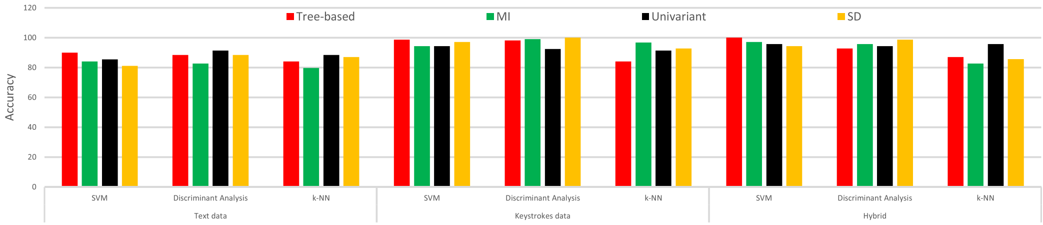

Fig. 5. Accuracy achieved on DEKT-345 × 2 database, two affective states.

Table 7. Summary of Classifiers and Features Selection Techniques Used in Best Results Achieved for Five and Binary Cslasses

Classes	Model	Classifier	Feature Selection Technique	Accuracy
Five	Text	SVM	Tree Based	86.53%
	Keystroke	SVM	Tree Based	65.22%
	Hybrid	Discriminant Analysis	MI Gain	86.95%
Two	Text	Discriminant Analysis	Statistical Dependency	100%
	Keystroke	Discriminant Analysis	Univariate	91.30%
	Hybrid	SVM	Tree Based	100%

selection technique and discriminant analysis with MI obtained 86.95% accuracy, while *k*-NN had only 55.07% accuracy.

5.2 Results on DEKT-345 × 2 for Dimensional Model of Emotions, Pleasure vs. Displeasure

Experiments for the two-dimensional model of emotions are described in this section. Accuracies achieved on keystrokes data for the two affective states are given in Figure 5. SVM has 89.85% accuracy with a tree-based feature selection technique, the discriminant analysis yielded 91.30% accuracy with univariate feature selection technique, and *k*-NN gave 88.41% accuracy with univariate feature selection technique method. Accuracies achieved by the CI models on text data for two affective states of emotions are given in Figure 5. SVM classifier with tree-based feature selection technique had 98.55% maximum accuracy on this model, the discriminant analysis gave 100% accuracy with statistical dependency-based feature selection technique, and *k*-NN yielded 96.65% accuracy using MI for feature selection. The hybrid approach for two dimensions of emotions is evaluated using the same feature selection techniques and classifiers. Accuracies achieved on the hybrid model are given in Figure 5. It has 100% accuracy using SVM classifier with tree-based feature selection technique, 98.55% accuracy using discriminant analysis classifier, and statistical dependency feature selection technique and 95.65% accuracy using *k*-NN classifier and univariate feature selection technique method.

From the above results, it can be seen that the tree-based feature selection technique and SVM classifier has the best accuracy on text data (86.53%) and keystrokes data (65.22%). MI technique and discriminant analysis had 86.95% accuracy on the hybrid data. Based on the two-dimensional emotion model, the statistical dependency technique and discriminant analysis had the best accuracy on text data (i.e., 100%), univariate feature selection technique and discriminant analysis had the best accuracy (i.e., 91.30%) on keystrokes data. Whereas, tree-based feature selection technique and SVM performed better on hybrid data and achieved 100% as the best accuracy. The same is represented in Table 7.

Table 8. Comparison of Previous Models' Results with the Proposed Models on DEKT-345 × 2

Works	Data	Approach	Accuracy	Accuracy (Proposed Approach)
Shikder et al. [12]	Keystroke dynamics (5 class classification)	Statistical Features	21.74%	65.22%
Tripathi et al. [21]	Text (5 class classification)	Conv1D	47.83%	86.53%
Tripathi et al. [21]	Text (5 class classification)	LSTM	48.20%	86.53%
Alhuzali et al. [22]	Text	GRU	85.71%	100%
Aguado et al. [51]	Keystroke dynamics (5 class classification)	ANN	25.71%	65.22%
Aguado et al. [51]	Keystroke dynamics (binary classification)	ANN	82.86%	91.30%
Aguado et al. [51]	Text (5 class classification)	ANN	60%	86.53%
Aguado et al. [51]	Text (binary classification)	ANN	83.23%	100%

5.3 Comparison of Prior Work with the Proposed Solution on DEKT-345 × 2 Data

The proposed approach is compared with three state-of-the-art methods for emotion identification. These include the proposal by Shikder et al. [12], Tripathi et al. [21], Alhuzali et al. [22], and Aguado et al. [51]. The proposal by Shikder et al. [12] is based on mouse and keystrokes usage and they worked for five emotional states, namely happiness, inspiration, sympathy, disgust, and fear. They used existing classifiers, namely, *k*-NN, KStar, random committee, and random forest. They also propose a classifier bounded *k*-means clustering method of the task at hand. Their proposed classifier works equally with other classifiers in case of less dominant emotions and consumes less classification time. Their approach of the existing classifiers is used here for comparison since the focus of this work is also to obtain better classification performance. A comparison between the proposed approach and Shikder et al. is shown in Table 8. Alhuzali et al. [22] introduce a new dataset of Modern Standard and Dialectal Arabic emotion detection based on eight basic emotion types (introduced by Robert Plutchik). They implemented a deep-gated recurrent neural network (GRU) and compared its performance with SVM. A comparison of the proposed approach and Alhuzali et al. is also shown in Table 8. On the keystrokes data, the solution presented in [12] yielded 21.74% accuracy while the present proposal gives 65.22% accuracy. For the text data, the approach in [22] yielded 85.71% maximum accuracy for a happy class, whereas the proposed approach here gives 100% accuracy on text for the surprise class. The approach used in Tripathi et al. [21] for emotion recognition is based on deep learning models. They use Conv1D and LSTM models on the IEMOCAP dataset. Here, for the sake of fairer comparison, the architecture used in [40] is applied for text data of the DEKT-345 × 2 dataset. Tripathi et al. used LSTM and Conv1D models for text data. Results obtained by their approach are shown in Table 8. It can be seen from the results that their Conv1D model yields 47.83% accuracy. Whereas, the LSTM model has 48.20% accuracy. Therefore, it can be seen that the present work performs better than the three state-of-the-art methods in identifying the human emotion from the small text (DEKT-345 × 2 dataset).

As the results show, the proposed framework yields 86.53% accuracy on text data for the five classes of emotions. Other than the existing experiments, an additional experiment is performed to compare the proposed approach with Aguado et al. [51]. Aguado et al. [51] recently proposed a method for the detection of sentiment and stress levels of users navigating a social network site. They developed novel analyzer agents that are integrated into the existing multi-agent systems. They use text data and keystroke dynamics and utilized this dataset to train **Artificial Neural Networks** (**ANNs**) representing the agents. Architecture of their ANN consists of the layer in the sequence of dense layer, dropout layer, dense layer, dropout layer, dense layer, and sentiment classification.

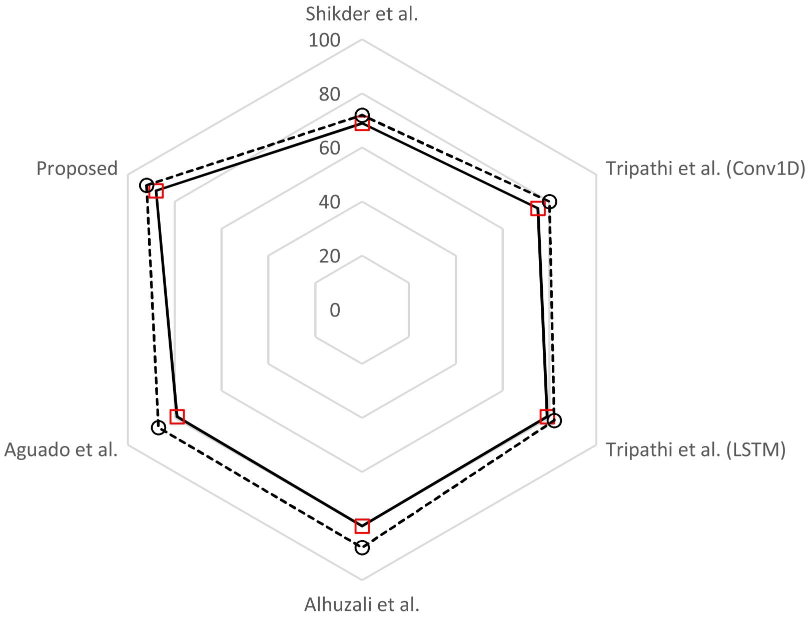

Fig. 6. Accuracy achieved on Enron8715 and Emotion616 datasets.

For a fairer comparison, the procedure of Aguado et al. is applied on our text and keystroke dynamics dataset and the obtained results are given in Table 8. Results show that 25.71% accuracy is achieved using keystrokes data for 5 class problem and 82.86% accuracy for binary class problem using Aguado et al. [51] method. Reason underlying these results can be the limited size of training dataset. While on text data, 60% accuracy is achieved for 5 class problem and 83.23% for binary classification.

5.4 Experiment with Additional Data

The present work performs simulations on the dataset specifically collected for the current study. However, it will be interesting to see the framework's performance on some previously available datasets with the same class labels. There are very limited such datasets publically available with the same class labeling as utilized in the present work [54, 55]. For this experiment, Enron8715 and Emotion616 datasets are obtained as mentioned in [32]. The Enron8715 is a labelled data of e-mails which is manually annotated utilizing the majority vote of multiple annotators. Each e-mail is given one of the six labels, i.e., neutral, happy, sad, angry, positively surprised, and negatively surprised. The dataset has 1,000 instances. The Emotion616 dataset also has the same number and type of classes as in Enron8715, but it is specifically collected from 61 volunteers by inducing emotion through stimuli. The dataset has 339 instances. Figure 6 shows the results of this experiment by executing the present framework and the five competing methods on Enron8715 and Emotion616 datasets. The obtained results show two major trends worth noticing. The first one is that the accuracies of all the methods including the proposed one have improved on Enron8715 and Emotion616 datasets as compared to their performance on DEKT-345 × 2 data. The primary reason for this is the availability of textual data only in the Enron8715 and Emotion616 datasets. Whereas, the DEKT-345 × 2 data also has keystrokes information. The second trend in this experiment is that better performance of almost all methods on Emotion616 as compared to the Enron8715 dataset. The basic reason for this is the presence of actual emotions in Emotion616 as compared to Enron8715, where labelling is done on free running e-mail text.

6 CONCLUSION

This work presented a novel text-keystrokes database, DEKT-345 × 2 using the English language text which contains text and keystrokes dynamics of five basic affective states. Utility of the database is shown by implementing baseline emotion recognition experiments on text and keystrokes using CI techniques. This work aimed to identify emotions captured in a real-world scenario while participants are not forced to write their feelings on a specific topic and unprompted to pen a predefined script. The participants were asked to express their feelings in maximum of 150 words, as the focus here was to detect emotions from a short text. Obtained results showed that the sum, mean, and standard deviation values of time duration of keys have high scores than others, while the frequency of space key and backspace key (rate of error) have scored higher than remaining frequency-based features, however, they score lesser than sum, mean, and keys' time standard deviation. Emotion detection in this work from text data was based on **Term Frequency–Inverse Document Frequency** (**TFIDF**) of the data. This work extracted both unigrams and bigrams from raw text data. The proposed work designed keystrokes, text, and text-keystrokes models and evaluated them on the DEKT-345 × 2 database. In case of text data, ngrams-based features were considered. Whereas for keystrokes, most of the features were time-based attributes. Models were designed for five basic emotions (i.e., happy, sad, angry, surprise, and disgusting) and pleasure emotion dimension (pleasure and displeasure). These were based on preprocessing methods, features selection techniques, such as, tree-based, univariate feature selection, MI, and statistical dependency. Three classifiers were applied, namely, SVM, discriminant analysis, and k-Nearest Neighbor (k-NN). The obtained results showed SVM and discriminant analysis performed better and the features extracted from tree-based feature selection gave good results in most of the scenarios. For training 80% of data was used and testing was performed on 20% data. Training and validation were done for 20 different combinations of training and validation data while testing was performed on unseen data. This work presented a real-life and non-acted text-keystrokes dataset, DEKT-345 × 2 collected from 69 subjects aged between 20 to 25 years. Experiments on the data showed that hybrid model and text-based model achieved higher accuracy. While experiments on keystrokes also performed well as compared to the past works. This proposal was also compared with four state-of-the-art techniques for the same task, where the present proposal performed better.

REFERENCES

[1] J. Liu, T. Li, P. Xie, S. Du, F. Teng, and X. Yang. 2020. Urban big data fusion based on deep learning: An overview. *Information Fusion* 53 (2020), 123–133.

[2] S. Ramírez-Gallego, A. Fernández, S. García, M. Chen, and F. Herrera. 2018. Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce. *Information Fusion* 42 (2018), 51–61.

[3] M. N. Bechtoldt, S. Rohrmann, I. E. De Pater, and B. Beersma. 2011. The primacy of perceiving: Emotion recognition buffers negative effects of emotional labor. *Journal of Applied Psychology* 96, 5 (2011), 1087.

[4] M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino. 2019. Human emotion recognition using deep belief network architecture. *Information Fusion* 51 (2019), 10–18.

[5] P. Ekman. 1999. Basic emotions. *Handbook of Cognition and Emotion.* John Wiley & Sons, USA, 45–60.

[6] R. Abaalkhail, B. Guthier, R. Alharthi, and A. El Saddik. 2018. Survey on ontologies for affective states and their influences. *Semantic Web* 9, 4 (2018), 441–458.

[7] R. Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist* 89, 4 (2001), 344–350.

[8] Donald L. Nathanson. 1992. *Shame and Pride: Affect, Sex, and the Birth of the Self.* W.W. Norton, New York.

[9] A. Valdivia, M. V. Luzón, E. Cambria, and F. Herrera. 2018. Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion* 44 (2018), 126–135.

[10] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera. 2018. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion* 44 (2018), 65–77.

[11] T. L. Nwe, S. W. Foo, and L. C. De Silva. 2003. Speech emotion recognition using hidden Markov models. *Speech Communication* 41, 4 (2003), 603–623.

[12] R. Shikder, S. Rahaman, F. Afroze, and A. A. Islam. 2017. Keystroke/mouse usage based emotion detection and user identification. In *Proceedings of the International Conference on Networking, Systems and Security (NSysS)*, 96–104.

[13] S. Salmeron-Majadas, R. S. Baker, O. C. Santos, and J. G. Boticario. 2018. A machine learning approach to leverage individual keyboard and mouse interaction behavior from multiple users in real-world learning scenarios. *IEEE Access* 6 (2018), 39154–39179.

[14] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31.

[15] Y. Gu, S. Chen, and I. Marsic. 2018. Deep multimodal learning for emotion recognition in spoken language. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.*

[16] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem. 2017. BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing* 8, 3 (2017), 300–313.

[17] R. L. Payne and C. Cooper (Eds.). 2003. *Emotions at Work: Theory, Research and Applications for Management.* John Wiley & Sons.

[18] J. R. Averill. 1983. Studies on anger and aggression: Implications for theories of emotion. *American Psychologist* 38, 11 (1983), 1145.

[19] R. Adolphs. 2017. How should neuroscience study emotions? By distinguishing emotion states, concepts, and experiences. *Social Cognitive and Affective Neuroscience* 12, 1 (2017), 24–31.

[20] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation.* 42, 4 (December 2008), 335–359.

[21] S. Tripathi and H. Beigi. 2018. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. Retrieved from https://arxiv.org/abs/1804.05788.

[22] H. Alhuzali, M. Abdul-Mageed, and L. Ungar. 2018. Enabling deep learning of emotion with first-person seed expressions. In *Proceedings of the 2nd Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media.* 25–35.

[23] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, and H. Prendinger. 2018. Decision support with text-based emotion recognition: Deep learning for affective computing. *Decision Support Systems* 115 (2018), 24–35.

[24] N. Sebe, I. Cohen, T. Gevers, and T.S. Huang. 2015. A multimodal approach to detect user's emotion. *Procedia Computer Science* 70, 1 (2015), 296–303.

[25] N. Gupta, M. Gilbert, and G. D. Fabbrizio. 2013. Emotion detection in email customer care. *Computational Intelligence* 29, 3 (2013), 489–505.

[26] Y. S. Seo and J. H. Huh. 2019. Automatic emotion-based music classification for supporting intelligent IoT applications. *Electronics* 8, 2 (2019), 164.

[27] S. Wolff, C. Stiglmayr, H. J. Bretz, C. H. Lammers, and A. Auckenthaler. 2007. Emotion identification and tension in female patients with borderline personality disorder. *British Journal of Clinical Psychology* 46, 3 (2007), 347–360.

[28] D. Stojanovski, G. Strezoski, G. Madjarov, and I. Dimitrovski. 2015. Emotion identification in FIFA world cup tweets using convolutional neural network. In *Proceedings of the 11th International Conference on Innovations in Information Technology.* 52–57.

[29] R. Subhashini and P. R. Niveditha. 2015. Analyzing and detecting employee's emotion for amelioration of organizations. *Procedia Computer Science* 48 (2015), 530–536.

[30] H. Liu, O. N. N. Fernando, and J. C. Rajapakse. 2018. Predicting affective states of programming using keyboard data and mouse behaviors. In *Proceedings of the 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 55–57.

[31] F. H. Leong. 2016. Fine-grained detection of programming students' frustration using keystrokes, mouse clicks and interaction logs. *Open Journal of Social Sciences* 4, 9 (2016), 9.

[32] Z. Halim, M. Waqar, and M. Tahir. 2020. A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. *Knowledge-Based Systems* 208 (2020), 106443.

[33] S. K. Khare, and V. Bajaj. 2020. Time-frequency representation and convolutional neural network-based emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 32, 7 (2020), 2901–2909.

[34] A. Kazemzadeh, S. Lee, and S. Narayanan. 2013. Fuzzy logic models for the meaning of emotion words. *IEEE Computational Intelligence Magazine* 8, 2 (2013), 34–49.

[35] M. S. Akhtar, A. Ekbal, and E. Cambria. 2020. How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble. *IEEE Computational Intelligence Magazine* 15, 1 (2020), 64–75.

[36] O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006. The eNTERFACE'05 audio-visual emotion database. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*. 8.

[37] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. 2005. A database of German emotional speech. In *Proceedings of the 9th European Conference on Speech Communication and Technology*.

[38] C. Epp, M. Lippold, and R. L. Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 715–724.

[39] A. N. H. Nahin, J. M. Alam, H. Mahmud, and K. Hasan. 2014. Identifying emotion by keystroke dynamics and text pattern analysis. *Behaviour & Information Technology* 33, 9 (2014), 987–996.

[40] S. Ghosh, N. Ganguly, B. Mitra, and P. De. 2017. Evaluating effectiveness of smartphone typing as an indicator of user emotion. In *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction*. 146–151.

[41] A. Kolakowska. 2015. Recognizing emotions on the basis of keystroke dynamics. In *Proceedings of the 8th International Conference on Human System Interaction*. 291–297.

[42] A. Kolakowska. 2016. Towards detecting programmers' stress on the basis of keystroke dynamics. In *Proceedings of the Federated Conference on Computer Science and Information Systems*. 1621–1626.

[43] S. Grover and A. Verma. 2016. Design for emotion detection of punjabi text using hybrid approach. In *Proceedings of the International Conference on Inventive Computation Technologies*, Vol. 2, 1–6.

[44] R. A. Calix, S. A. Mallepudi, B. Chen, and G. M. Knapp. 2010. Emotion recognition in text for 3-D facial expression rendering. *IEEE Transactions on Multimedia* 12, 6 (2010), 544–551.

[45] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S.S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* 42, 4 (December 2008), 335–359.

[46] J. M. Talarico, D. Berntsen, and D. C. Rubin. 2009. Positive emotions enhance recall of peripheral details. *Cognition and Emotion* 23, 2 (2009), 380–398.

[47] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li. 2018. Spatial–temporal recurrent neural network for emotion recognition. *IEEE Transactions on Cybernetics* 49, 3 (2018), 839–847.

[48] J. Li, S. Qiu, Y. Y. Shen, C. L. Liu, and H. He. 2019. Multisource transfer learning for cross-subject EEG emotion recognition." *IEEE Transactions on Cybernetics* 50, 7 (2019), 3281–3293.

[49] W. L. Zheng, W. Liu, Y. Lu, B. L. Lu, A. Cichocki. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics* 49, 3 (2018), 1110–1122.

[50] G. Pons and D. Masip. 2020. Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition. *IEEE Transactions on Cybernetics*. DOI: 10.1109/TCYB.2020.3036935.

[51] G. Aguado, V. Julián, A. García-Fornes, and A. Espinosa. 2020. Using keystroke dynamics in a multi-agent system for user guiding in online social networks. *Applied Sciences* 10, 11 (2020), 3754.

[52] C. Nguyen, T. Hassner, M. Seeger, and C. Archambeau. 2020. Leep: A new measure to evaluate transferability of learned representations. In *Proceedings of the International Conference on Machine Learning*. 7294–7305, PMLR.

[53] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al-Shamma, M. A. Fadhel, … and Y. Duan. 2021. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers* 13, 7, (2021), 1590.

[54] Z. Xu and S. Wang. 2021. Emotional attention detection and correlation exploration for image emotion distribution learning. *IEEE Transactions on Affective Computing*. DOI: 10.1109/TAFFC.2021.3071131

[55] Z. Xu, S. Wang, and C. Wang. 2020. Exploiting multi-emotion relations at feature and label levels for emotion tagging. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2955–2963.

Received January 2021; revised July 2021; accepted August 2021