IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 February 2024, accepted 11 May 2024, date of publication 15 May 2024, date of current version 24 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3401203

RESEARCH ARTICLE

Evaluating the Effect of Emotion Models on the Generalizability of Text Emotion Detection Systems

ALEJANDRO DE LEÓN LANGURÉ AND MAHDI ZAREEI, (Senior Member, IEEE)
Tecnológico de Monterrey, School of Engineering and Sciences, Zapopan, Jalisco 45201, Mexico
Corresponding author: Mahdi Zareei (m.zareei@tec.mx)

ABSTRACT Text emotion detection is a pivotal aspect of natural language processing, with wide-ranging applications involving human-computer interactions. Machine learning agents have been trained with supervised methods, thus relying on labeled datasets. However, the arbitrary selection of emotion models while labeling such datasets poses significant challenges in the performance and generalizability of the produced machine learning predictors, primarily when evaluated against unseen data, as it effectively introduces bias to the process. This study investigates the impact of emotion model selection on the efficacy of machine learning systems for text emotion detection. Eight labeled datasets were employed to train linear regression, feedforward neural network, and BERT-based deep learning models. Results demonstrated a notable decrease in accuracy when models trained on one dataset were tested on others, underscoring the inherent incompatibilities in labeling across datasets. To prove that the emotion model significantly impacts predictors' performance, we propose a standardized emotion label mapping utilizing James Russell's circumplex model of affect that turns the emotion model into a parameter rather than a fixed element. Cross-dataset testing with this shared emotion mapping yielded significant, non-negligible changes in accuracy (both improvement and degradation). This fact highlights the impact of the emotion model (traditionally arbitrarily selected) during machine learning training and performance, arguing that improvements in accuracy reported in related research literature might be due to differences in the used emotion model rather than the new algorithms introduced.

INDEX TERMS Affective computing, natural language processing, sentiment analysis, text emotion detection, text emotion recognition.

I. INTRODUCTION

In the burgeoning field of natural language processing (NLP), detecting and interpreting emotions in text represents a fundamental challenge with far-reaching implications [1], [2] [3], [4]. The ability to discern emotions accurately from textual data has transformative potential across various applications, including sentiment analysis, interactive chatbots, and tailored recommendation systems. At the heart of this challenge lies the selection and application of emotion models, which serve as foundational frameworks for labeling and interpreting the emotional content of text.

The diversity of emotion models used for labeling datasets poses a significant challenge in generalizing machine learning algorithms, hindering their ability to handle unseen data effectively. A text emotion detection model should demonstrate consistent performance across various scenarios in practical applications. However, due to the cost of producing labeled datasets, current research often relies on existing resources to propose enhancements in emotion prediction accuracy. Unfortunately, these datasets are labeled with inconsistent emotions, leading to hyper-optimized models for specific datasets prone to overfitting. This critical issue impedes the deployment of these models in real-world scenarios, underscoring the practical implications of our research.

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

Furthermore, current research often introduces new machine learning algorithms purportedly enhancing accuracy but typically sidesteps cross-dataset testing. Consequently, it can be argued that these studies only demonstrate the new algorithm's heightened overfitting to the used dataset rather than validating its generalization capabilities.

This research aims to highlight the influence of selecting an arbitrary emotion model on machine learning algorithms' performance and generalization capabilities in text emotion detection. We propose a novel emotion labeling approach centered around a universally recognized framework such as James Russell's circumplex model of affect. We demonstrate significant changes in ML models' performance by manipulating the emotion model while keeping all other elements the same. This underscores the novelty of our approach and the importance of considering emotion models as variable hyperparameters, an element often overlooked in current literature.

II. PRELIMINARIES

A. SENTIMENT AND EMOTIONS IN TEXT

Two fundamental domains within NLP for text encompass sentiment analysis (SA) and text emotion detection (TED) [5]. SA focuses on discerning whether data conveys positivity, negativity, or neutrality, which can be evaluated along a single dimension. In contrast, TED delves into identifying a spectrum of human emotion categories, including but not limited to anger, happiness, or sadness, necessitating a more intricate analytical approach. Central to TED is using an emotion model (EM) as a theoretical framework or guideline for understanding and classifying emotions within textual data. An EM encompasses a structured representation of emotion categories, facilitating the classification of emotions in text data. By leveraging predefined emotion categories, the EM enables the accurate identification of emotional states conveyed by the text, thereby enhancing the interpretability and utility of TED systems.

EMs can be divided into two primary types: categorical and dimensional [6]. Categorical models offer discrete classifications or labels for emotions. For instance, Robert Plutchik's wheel of emotions delineates eight primary emotions (joy, trust, fear, surprise, sadness, disgust, anger, and anticipation), which can be further combined to generate secondary emotions [7]. Similarly, Carroll Izard's differential emotions theory posits ten basic emotions, including interest, joy, surprise, anger, sadness, disgust, contempt, shame, guilt, and fear [8]. Meanwhile, Paul Ekman's six basic emotions model identifies anger, disgust, fear, happiness, sadness, and surprise [9].

In contrast, dimensional EMs do not lend themselves to discrete representation. Instead, they are positioned along a continuum defined by axes. Dimensional EMs, such as the affective circumplex model by James Russell, conceptualize emotions as points on a two-dimensional space defined by valence (ranging from positive to negative) and arousal (from low to high) [10]. Klaus Scherer's three-dimensional model of emotion incorporates valence, arousal, and a third dimension denoting the perceived degree of control over emotion [11]. James Gross's component process model proposes that emotions stem from basic component processes, including affective responses, cognitive appraisals, and physiological changes [12]. Influenced by contextual factors, the appraisal component shapes individuals' emotional experiences based on their interactions with their environment and past encounters. Additionally, the pleasure-arousal-dominance (PAD) model, developed by Mehrabian and Russell, augments the circumplex model by introducing a third dimension, dominance, representing the perceived degree of control or power associated with an emotion [13].

B. EMOTIONS AS NUMERICAL VALUES

In Russell's circumplex model of affect, valence represents an emotion's positive or negative nature, ranging from pleasant to unpleasant. Emotions with positive valence are typically associated with happiness, joy, and contentment, while those with negative valence include sadness, anger, and fear. Arousal, on the other hand, reflects an emotion's intensity or activation level, ranging from low to high. Emotions with low arousal are calm and relaxed, while those with high arousal are intense and stimulating.

This model represents emotions as points in a two-dimensional space defined by valence and arousal axes. They can, therefore, be assigned numerical pairs of values analogous to cartesian coordinates. This representation allows for a comprehensive understanding of various emotional states by categorizing them based on their positions within this space. For example, emotions like excitement and euphoria would be located in the high arousal, positive valence quadrant, while emotions like depression and fatigue would be in the low arousal, negative valence quadrant. By mapping emotions this way, Russell's model provides a structured framework for analyzing and interpreting emotional experiences, facilitating research in psychology, neuroscience, and TED, among others.

C. MACHINE LEARNING

Machine learning (ML) is a subfield of artificial intelligence (AI) that develops algorithms and models to enable computers to learn from data, make predictions, and make decisions [14]. Unlike traditional rule-based programming, where explicit instructions are provided for solving problems, ML systems learn patterns and relationships directly from data by identifying underlying structures and adjusting their parameters.

ML algorithms can be broadly categorized into supervised, unsupervised, and reinforcement learning paradigms. In supervised learning, the algorithm is trained on labeled data, where each example is associated with a target output. Unsupervised learning involves discovering patterns or structures in unlabeled data. In contrast, reinforcement learning relies on learning through trial and error, with the algorithm receiving feedback as rewards or penalties based on its actions.

One of ML's critical strengths is its ability to generalize from training data to make predictions or decisions on new, unseen data. This generalization capability enables ML models to adapt and perform well in diverse and complex scenarios.

One application of ML is TED, where algorithms are trained to recognize and classify emotions conveyed in text, such as those found in social media networks, internet blogs, review sites, or online newspapers. In this context, ML leverages an EM, which serves as the framework for categorizing emotions within the text. The process typically involves training the ML model using labeled datasets, where each text sample is associated with one or more emotion labels based on the chosen EM. During training, the algorithm learns to identify patterns and relationships between the features extracted from the text data and the corresponding emotion labels. This process often involves adjusting the model's parameters to optimize its performance in predicting the correct emotion labels for new, unseen text samples.

Once the model is trained, it undergoes testing and validation to assess its performance and generalization capabilities. Testing involves evaluating the model's accuracy and effectiveness in classifying emotions on a separate dataset not used during training. Validation ensures the model performs reliably across different datasets and scenarios, helping identify and address potential biases or limitations.

The present research will discuss three ML algorithms: linear regression, simple feedforward neural networks, and BERT.

1) LINEAR REGRESSION

In TED, linear regression is a predictive modeling technique to infer the emotional states associated with textual data [15]. It establishes a linear relationship between the features extracted from the text, such as word frequencies or TF-IDF scores, and the corresponding emotion labels. Through the learning process, the linear regression model attempts to discern underlying patterns and associations within the textual features to predict the emotional responses conveyed by the text. Upon completing the training phase, the model can generate predictions for new text samples based on their feature representations, assigning probabilities to different emotion categories.

2) FEEDFORWARD NEURAL NETWORKS

A feedforward neural network (FNN) is a fundamental type of artificial neural network (ANN) commonly used for various ML tasks, including TED. Unlike other neural network architectures, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), FNNs do not have feedback loops or cyclic connections between neurons. Instead, they consist of multiple layers of interconnected neurons, with information flowing strictly in one direction, from the input layer through one or more hidden layers to the output layer [16].

In the context of TED, an FNN is employed to learn the complex mappings between textual features extracted from the input data and the corresponding emotion labels. The architecture of an FNN typically includes an input layer, one or more hidden layers, and an output layer. Each neuron in the input layer represents a feature of the input data, such as word frequencies or TF-IDF scores. In contrast, neurons in the hidden layers perform computations and extract higher-level representations of the input data. The output layer produces predictions or classifications based on the learned representations, with each neuron corresponding to a different emotion category.

The training process of an FNN involves iteratively adjusting the weights and biases of the network's connections to minimize a predefined loss function, typically using an optimization algorithm such as stochastic gradient descent (SGD) or Adam. A FNN can be trained using a backpropagation algorithm, where gradients of the loss function concerning the network parameters are computed and used to update the parameters in the opposite direction of the gradient. During training, the FNN learns to map the input text features to the correct emotion labels, enabling it to make accurate predictions on new, unseen text samples.

3) BERT

In TED, using BERT (Bidirectional Encoder Representations from Transformers) leverages pre-trained transformer-based language models. BERT, a deep learning model developed by Google, captures contextual information from textual data by considering both left and right contexts bidirectionally. Unlike traditional models, BERT does not require handcrafted features or prior feature engineering, as it learns contextual representations directly from the text. BERT models can be fine-tuned for sequence classification tasks, where they learn to predict emotion labels from input text data [17]. The training begins with initializing the BERT tokenizer and model, utilizing a pre-trained model for text tokenization and sequence classification. Texts and corresponding emotion labels are extracted from the dataset, and labels are encoded for numerical representation. Texts are then tokenized and encoded using the BERT tokenizer, incorporating special tokens and truncation/padding to ensure uniform input length.

During the training loop, a BERT model can be improved using the Adam optimizer with a cross-entropy loss function to minimize the discrepancy between predicted and actual emotion labels. The model is trained over multiple epochs, with gradients calculated and updated using backpropagation. Evaluation of the validation set involves making predictions using the trained model and computing evaluation metrics such as accuracy, precision, recall, and F1 score.

D. PYTHON IMPLEMENTATION

PyTorch and scikit-learn are two prominent Python libraries widely used in ML and data science. PyTorch is an open-source ML framework primarily known for its flexibility and dynamic computational graph construction [18]. It provides a platform for building and training NNs, offering extensive support for deep learning tasks such as image classification, NLP, and reinforcement learning. PyTorch's popularity stems from its intuitive interface, which allows researchers and practitioners to experiment with complex models and algorithms efficiently.

On the other hand, scikit-learn is a comprehensive ML library built on top of Python's scientific computing stack [19], including NumPy, SciPy, and Matplotlib. Unlike PyTorch, scikit-learn focuses on traditional ML algorithms, providing implementations for various supervised and unsupervised learning techniques such as classification, regression, clustering, and dimensionality reduction.

The experiments for this research paper were implemented in Python using the PyTorch and scikit-learn libraries.

E. MEASURING PERFORMANCE IN ML ALGORITHMS

TED, fundamentally a classification task, prioritizes prediction and accuracy as success metrics [20], [21]. While precision holds value, this research does not center on constructing a novel, finely-tuned ML algorithm. Instead, we investigate the impact of EM selection on prediction accuracy. Therefore, precision takes a second plane. To isolate this effect, we intentionally employed three established algorithms (linear regression, FNN, and BERT) without extensive optimization.

Previous research in TED demonstrates that peer-accepted performance improvements typically range from 5% upwards [22], [23], [24], [25]. While lacking a universally standardized definition of "significant change" within this field, our study aligns with the benchmarks established by current research and state-of-the-art findings. By adhering to these recognized improvement measures, we provide a contextualized basis for evaluating the impact of our research.

F. HYPERPARAMETERS

In ML, hyperparameters are preconditions or configurations set before the learning process begins and not directly learned from the data during training. They control aspects of the learning process and model architecture, such as the learning rate, the number of hidden layers in a NN, or the choice of kernel in a support vector machine. Unlike model parameters, which are learned from the data, hyperparameters must be chosen beforehand by the practitioner based on intuition, experimentation, or domain knowledge [26].

Selecting appropriate hyperparameters can improve model performance, generalization, and faster convergence during training. Conversely, poorly chosen hyperparameters can result in suboptimal performance, overfitting, or underfitting of the model. Thus, effectively tuning hyperparameters is essential for obtaining the best possible performance from an ML model.

In linear regression, common hyperparameters include the learning rate (for gradient descent-based optimization algorithms), the regularization parameter (e.g., L1 or L2 regularization strength), and the choice of optimization algorithm. In FNN, hyperparameters include the number of hidden layers, the number of neurons in each layer, the activation function used in each layer, the learning rate, and the batch size. For BERT-based classifiers or other transformer models, hyperparameters include the learning rate, the number of layers, the number of attention heads, the sequence length, the batch size, and the choice of pretraining checkpoint.

III. EXPERIMENTS

A. PROBLEM STATEMENT

Current research on TED has demonstrated the remarkable capability of state-of-the-art algorithms to achieve prediction accuracies exceeding 90% [27]. However, while considerable emphasis has been placed on algorithm selection and hyperparameter tuning, other critical factors, such as the choice of EM, have often been overlooked. In a significant portion of literature and survey papers, approximately 77% fail to consider the influence of EMs when enumerating factors affecting model performance [28].

From the ML research perspective, TED is typically approached as a classification task, with emotions serving as the classifiers based on the chosen EM. Various EMs have been utilized as the foundation for TED research without explicit justification for their selection. Moreover, the proliferation of classifiers within these EMs contributes to algorithmic complexity, potentially obscuring the true determinants of model performance. Thus, disparities in performance across trained models may be attributed not only to computational limitations during training but also to challenges in generalization owing to the differing number of classifiers [29].

Given these considerations, the absence of a standardized number of classifiers, stemming from the utilization of diverse EMs, hinders the establishment of baseline computational requirements necessary for optimal model fitting. Consequently, discrepancies in accuracy performance may partly result from disparities in hardware resources. Moreover, existing research predominantly focuses on training and validating model performance within closed datasets, neglecting cross-dataset testing on unseen data. The rationale often cited for this omission is the incorporation of k-fold cross-validation during training, deeming external dataset testing unnecessary [30], and research has been proven to contribute to the body of knowledge using this methodology [31].

Furthermore, the incomparability of datasets in terms of quality, distribution, and emotional categories poses a significant challenge. Each dataset is tailored to specific

TABLE 1. Overview of included labeled datasets.

id	Name	Description	Examples	Emotions	Ref
1	Emotions Detection from Text	It contains labels for the emotional content (such as happiness, sadness, and anger) of texts.	40,000	13	[32]
2	Emotions dataset for NLP	List of text documents with emotion flag split into train, test, and validation sets for machine learning modeling.	20,000	6	[33]
3	Emotion	A dataset of English Twitter messages with six basic emotions	416,810	6	[34]
4	ISEAR	Data set containing reports on 7 emotions, each by close to 3000 respondents in 37 countries on all 5 continents.	7,666	7	[35]
5	Google GoEmotions	Human-annotated dataset of Reddit comments extracted from popular English-language subreddits and labeled with 27 emotion categories.	58,009	28	[36]
6	Affective Text Dataset (SemEval-2007 competition)	Data set consisting of news headlines drawn from major newspapers such as New York Times, CNN, and BBC News, as well as from the Google News search engine.	1,250	6	[37]
7	Emotion-Stimulus dataset in NLP	Sentences with emotion and a cause of that emotion.	2,414	7	[38]
8	WASSA-2017 Shared Task on Emotion Intensity (EmoInt)	Dataset consisting on emotion annotated tweets, including a degree of intensity.	3,960	4	[39]

objectives or domains, rendering models trained on one dataset incompatible with others and performance measurement practically infeasible due to variations in emotion tags. Thus, the question of the arbitrary EM selection's impact on a model's performance arises.

B. RESEARCH QUESTION

How does the arbitrary selection of an EM affect the performance and generalizability of ML models for TED, and how can emotion labels from different models be standardized to improve compatibility and generalization across diverse datasets?

C. OBJECTIVES

This study investigates the impact of EM selection on the performance and generalizability of ML models for TED. A method is proposed to establish equivalence between emotion tags across models to enable meaningful comparisons between models trained on datasets with different emotion label schemes to establish equivalence between emotion tags across models. ML models will then be trained on each dataset using standardized labels derived from the original EM. The performance of these dataset-specific models will be compared to models trained on optimized, shared emotion labels mapped across datasets. Model performance will be evaluated both on the dataset used for training and by cross-testing on datasets with different underlying EMs. This cross-dataset testing is intended to assess the generalization ability of models trained under different conditions. By comparing performance within and across datasets, the study seeks to reveal patterns in how EM choice affects model accuracy and generalizability in TED tasks. The goal is to gain insight into how the selection of EMs impacts ML model development for real-world applications.

D. HYPOTHESIS

1) MAIN HYPOTHESIS

The selection of the emotion model impacts the performance and generalization of machine learning models for text emotion detection of at least 5% change in accuracy.

2) NEGATIVE HYPOTHESIS

The choice of emotion model has no substantial effect (less than 5% change) on the performance and generalizability of machine learning models for text emotion detection. Any observed differences can be attributed to variations in data quality or model architecture.

E. METHODOLOGY

1) DATA COLLECTION

Eight publicly available text datasets, each annotated with emotion labels according to distinct EMs, were obtained for this study, as illustrated in Table 1. During the initial acquisition phase, datasets were downloaded, decompressed, and consolidated into a shared directory. A Python program was developed to systematically read, parse, and format the disparate datasets into a unified SQLite3 database to facilitate subsequent analysis. This centralized database stored all collected text samples within a single table, streamlining data access and enabling cross-dataset comparisons.

The selection criteria for these datasets include general public availability, previous usage of peer-accepted research papers, and compatibility with the tools we used during experimentation.

At this point, we applied only basic text pre-processing, as shown in the following piece of code, and the regex responsible for text cleaning.

```
def clean_text(text):
    text = re.sub(r’\s+’, ’␣’, text)
    return re.sub(r’[^\w\s@#]’, ’’, text)
```

2) EMOTION LABEL STANDARDIZATION

The table containing all the datasets was defined with the following columns:

- id: Just a numerical consecutive.
- source_id: The id assigned by the dataset. Some of the datasets provided an ID for the text. For example, those datasets containing tweets provided the tweet ID.
- dataset_id: Each dataset was named from one to eight, as "DatasetFourLoader" for instance.
- text: The actual text as labeled example.
- source_emotion: The emotion or label as assigned by the source.
- shared_emotion: An emotion label assigned per the first EM mapping dictionary defined in this research.
- quadrant_emotion: An emotion label assigned per the second EM mapping dictionary defined in this research.

The initial hypothesis posits that arbitrarily selecting an EM significantly influences a TED model's performance. If valid, this suggests that EMs should be considered hyperparameters and externalized to allow optimization independently of the ML algorithm. We have implemented this EM optimization process as EM mappings, making it possible to enrich the original EM and produce a new one. This study introduces two EM mappings: "shared emotions" and "quadrant emotions."

Since each dataset was labeled independently, there is no formal methodology to create a unified set of labels for comparison despite some coincidental overlaps. As depicted in Table 1, the number of distinct labels varies from 4 to 28 across datasets. The first mapping, "shared emotions," aims to mitigate label disparities and reduce their overall count, thus minimizing the number of classifiers required.

It is imperative to highlight that this "shared emotions" mapping, like any hyperparameter, remains arbitrary but can be tailored and fine-tuned to enhance performance. In essence, this mapping was not hardcoded in the experiments, allowing practitioners the flexibility to modify it according to their needs. For this research, the "shared emotions" mapping reduced the possible labels to 10 emotions and a "neutral" one, as presented in Table 2.

The second mapping is denoted as "quadrant emotion," which draws from Russell's circumplex model of affect, delineating quadrants based on valence and arousal. Each distinct emotion tag, as labeled independently by respective datasets, is assigned arbitrary valence and arousal values within the range of -1 to 1, as presented in Table 3. Consequently, this mapping yields a Cartesian plane wherein emotions are positioned within one of four quadrants (or quadrant 0 in the case of neutral emotion). The designation for the "quadrant emotion" corresponds to the specific quadrant on the Cartesian plane where the emotion is situated (Q0, Q1, Q2, Q3, Q4).

Once more, as with the first mapping, this equivalence is provided as a Python dictionary and considered a hyperparameter. Therefore, practitioners can change and tune these values and see how performance might improve.

TABLE 2. "Shared emotions" mapping.

ID	Source Emotion	Shared Emotion
1	happiness	happy
2	happy	happy
3	fun	happy
4	joy	happy
5	disappointment	sad
6	grief	sad
7	remorse	sad
8	sad	sad
9	sadness	sad
10	anger	anger
11	hate	anger
12	fear	fear
13	guilt	fear
14	surprise	surprise
15	realization	surprise
16	annoyance	disgust
17	boredom	disgust
18	disapproval	disgust
19	disgust	disgust
20	caring	love
21	desire	love
22	gratitude	love
23	love	love
24	admiration	excitement
25	amusement	excitement
26	curiosity	excitement
27	enthusiasm	excitement
28	excitement	excitement
29	confusion	anxiety
30	embarrassment	anxiety
31	nervousness	anxiety
32	shame	anxiety
33	worry	anxiety
34	approval	contentment
35	optimism	contentment
36	pride	contentment
37	relief	contentment
38	empty	neutral
39	neutral	neutral

The number of labels for each dataset in their original EM, shared, and quadrant mappings can be seen in Table 4.

We wrote a Python program to implement both emotion mappings and store the results in the same database. An excerpt of the resulting database can be seen in Table 5. Both emotion mappings are provided as Python dictionaries.

The data collection and processing flow is represented in Figure 1.

3) MODEL TRAINING AND EVALUATION

We selected three distinct algorithms for our experimental setup: linear regression, FNN, and BERT. Each algorithm was employed to train models corresponding to different datasets and emotion mapping combinations. Specifically, for dataset one, three models were trained using linear regression: one with the original emotion labels, another with the shared emotion mapping, and a third with the quadrant emotion mapping. This exact procedure was repeated for dataset one across the FNN and BERT implementations, and this process was repeated with all datasets individually.

Consequently, our experimentation trained 72 distinct models (including all eight datasets, three algorithms,

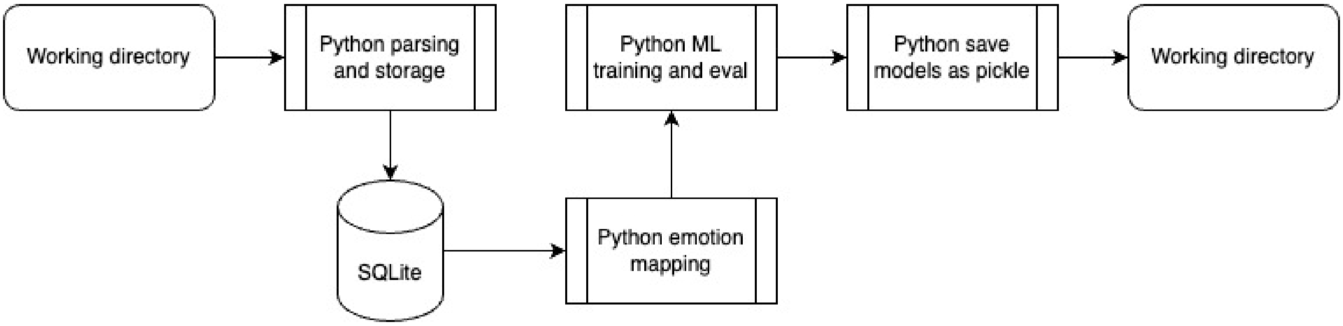
Working directory
Python parsing and storage
SQLite
Python emotion mapping
Python ML training and eval
Python save models as pickle
Working directory

FIGURE 1. Conceptual representation of the workflow during experimentation.

TABLE 3. "Quadrant emotions" mapping.

Source Emotion	Valence	Arousal
admiration	0.9	0.3
amusement	0.8	0.7
anger	-0.9	0.9
annoyance	-0.7	0.7
approval	0.8	0.3
boredom	-0.8	0.3
caring	0.8	0.3
confusion	-0.7	0.7
curiosity	0.7	0.7
desire	0.7	0.8
disappointment	-0.8	0.7
disapproval	-0.8	0.7
disgust	-0.9	0.8
embarrassment	-0.7	0.7
empty	-0.6	0.3
enthusiasm	0.8	0.8
excitement	0.9	0.9
fear	-0.9	0.9
fun	0.8	0.7
gratitude	0.9	0.3
grief	-0.9	0.7
guilt	-0.8	0.7
happiness	0.9	0.8
happy	0.9	0.8
hate	-0.9	0.9
joy	0.9	0.9
love	0.9	0.8
nervousness	-0.7	0.8
neutral	0.0	0.0
optimism	0.9	0.7
pride	0.8	0.7
realization	0.8	0.7
relief	0.8	0.3
remorse	-0.8	0.7
sad	-0.9	0.7
sadness	-0.9	0.7
shame	-0.7	0.7
surprise	0.8	0.9
worry	-0.8	0.8

TABLE 4. Comparison of labels after emotion mapping, ordered by dataset id.

id	Labels (source)	Labels (shared)	Labels (quad)
1	13	10	3
2	6	6	2
3	6	6	2
4	7	6	2
5	28	11	3
6	6	6	2
7	7	7	2
8	4	4	2

and three emotion mappings). As each model represents a unique combination of algorithm, dataset, and emotion mapping, this setup comprehensively evaluates the effectiveness of various approaches in handling emotion classification tasks.

The accuracy results of the trained models are presented in Tables 6, 7, and 8. It is important to emphasize that the primary focus of this research is not on algorithm optimization or achieving the highest possible accuracy scores. Instead, the objective is solely to assess the influence of the EM as a hyperparameter. As a result, the accuracy scores obtained may appear modest compared to state-of-the-art benchmarks. However, these scores serve as valuable indicators of the relative performance of the models under different EM configurations, shedding light on the impact of these hyperparameters on the overall effectiveness of emotion classification tasks.

The tables contain columns denominated "Delta," which represent the changes in accuracy related to the emotion mapping applied.

After training all the models with their corresponding datasets, the subsequent experimentation phase involves utilizing each trained model to assess the remaining datasets. For instance, we employ the linear regression model trained with dataset one and the corresponding EM to forecast labels in the other datasets. We randomly selected 1,000 examples from each dataset to conduct this evaluation and recorded the model's performance. This process was iterated three times (with different random seeds) to generate an average accuracy score for predicting unseen data.

This procedure was replicated across all datasets, algorithms, and emotion mappings. Each iteration involved sampling 1,000 random instances from the remaining datasets, repeating the process three times, and reporting the average performance.

Tables 9, 10, and 11 present the findings from cross-dataset testing. These tables are organized based on the dataset ID, with the first column delineating the trained

TABLE 5. Excerpt of the resulting database and the implemented emotion mappings.

ID	Source id	Dataset id	Text	Source	Shared	Quadrant	Valence	Arousal
1	1956967666	DatasetOneLoader	Layin n bed with a headache ughhhhwaitin on your call	sadness	sad	Q4	-0.9	0.7
2	1956967696	DatasetOneLoader	Funeral ceremony gloomy friday	sadness	sad	Q4	-0.9	0.7
3	1956967789	DatasetOneLoader	wants to hang out with friends SOON	enthusiasm	excitement	Q1	0.8	0.8

TABLE 6. Accuracy comparison of linear regression models, ordered by dataset id, expressed in percentages.

id	Original EM	Shared	Delta	Quadrant	Delta
1	35.03	38.01	2.99	59.51	24.49
2	85.33	85.33	0.00	94.33	9.00
3	89.35	89.35	0.00	96.84	7.50
4	58.74	58.80	0.07	90.35	**31.62**
5	39.74	44.32	**4.57**	61.24	21.50
6	44.00	44.00	0.00	69.20	25.20
7	83.44	83.44	0.00	82.19	-1.24
8	81.94	81.94	0.00	83.84	1.89

TABLE 7. Accuracy comparison of FNN models, ordered by dataset id, expressed in percentages.

id	Original EM	Shared	Delta	Quadrant	Delta
1	23.69	25.54	1.85	40.31	16.63
2	73.93	74.25	0.33	93.90	19.98
3	84.28	84.47	0.19	94.23	9.95
4	48.17	49.87	1.69	71.51	23.34
5	28.65	33.42	**4.77**	46.19	17.54
6	50.40	50.80	0.40	45.60	-4.80
7	56.94	57.14	0.21	93.17	**36.23**
8	72.35	72.73	0.38	87.25	14.90

TABLE 8. Accuracy comparison of BERT models, ordered by dataset id, expressed in percentages.

id	Original EM	Shared	Delta	Quadrant	Delta
1	37.41	40.31	2.90	62.43	25.01
2	93.25	93.90	0.65	98.13	4.88
3	94.17	94.23	0.05	98.87	4.69
4	71.06	71.51	0.46	97.07	26.01
5	41.76	46.19	**4.42**	63.43	21.67
6	43.20	45.60	2.40	80.00	**36.80**
7	93.79	93.17	-0.62	98.76	4.97
8	88.89	87.25	-1.64	92.05	3.16

model's original accuracy, the second column its accuracy during cross-dataset testing, and the third column showing the corresponding difference in performance. Subsequent columns in each table display analogous data but utilize shared and quadrant emotion mapping.

Table 9 pertains to the outcomes of cross-dataset testing conducted with the linear regression model. In contrast, Table 10 delineates the results obtained from the FNN model. Finally, Table 11 presents and compares the outcomes associated with the BERT models.

IV. DISCUSSION

The eight datasets utilized in this study were developed independently by various researchers, lacking compatibility considerations. Each dataset exhibited distinct characteristics, including unique label sets (EM), variations in class representation (class imbalances), and diverse domains. Based on their origins, it can be inferred that the individuals involved in labeling and the original authors of the texts represent disparate demographics, each with distinct perspectives on emotional expression.

It is essential to keep in mind that the objective of these experiments was not to propose a better algorithm for TED (with better performance) but to observe the changes in performance, as low as it might be initially, derived from changing only the selected EM for training while maintaining the same algorithm, same dataset, and same hyper-parameters.

Regarding the initial emotion mapping, referred to as shared emotions, no significant changes (less than 5%) in prediction accuracy or performance were observed while maintaining consistent datasets and algorithms, as can be seen in the first delta column in tables 6, 7, and 8. The most significant performance variations were observed in models trained with dataset 5, specifically Google GoEmotions, but still below the 5% mark. This discrepancy can be attributed to significant alterations in this dataset labels, resulting from the emotion mapping and label reduction process, reducing the label count from 28 to only eleven shared emotions (table 4).

An important observation stemming from the minimal impact of the initial emotion mapping adjustment is the consistency it implies in the algorithms' behavior. This consistency indicates that the algorithms remained stable without significantly introducing imbalance or skewness resulting from the modification in emotion mapping.

In the extreme scenario of maximum label reduction (as the quadrant emotion only allows for up to 5 labels), the second delta column in tables 6, 7, and 8 demonstrates a remarkable increase in prediction accuracy. The available labels were minimized under this second emotion mapping, while the algorithm and datasets remained unchanged.

The findings underscore the sensitivity of ML models to label modifications, suggesting that label mapping serves as a critical hyperparameter in model optimization. However, it is essential to exercise caution when interpreting results in the context of extreme label customization, as those in the second delta column from tables 6, 7, and 8, since they may lead to exaggerated or unrealistic performance outcomes. Effectively, as with any other hyper-parameter, EM manipulation can lead to model overfitting.

TABLE 9. Accuracy comparison of cross-dataset testing for linear regression models, ordered by dataset id, expressed in percentages.

id	Original EM	Original EM	Original	Shared EM	Shared Delta	Quad EM	Quad Delta
	accuracy	cross-testing	Delta	cross-testing		cross-testing	
1	35.03	9.10	-25.93	12.33	-22.69	57.40	22.38
2	85.33	55.13	-30.19	55.20	-30.13	76.33	-8.99
3	89.35	12.00	-77.35	16.17	**-73.18**	52.70	-36.65
4	58.74	39.53	-19.20	12.03	-46.70	43.83	-14.90
5	39.74	6.47	-33.28	12.40	-27.34	64.30	24.56
6	44.00	21.17	-22.83	22.53	-21.47	45.47	1.47
7	83.44	3.03	**-80.40**	24.93	-58.50	47.13	-36.30
8	81.94	17.13	-64.81	17.93	-64.01	45.17	**-36.78**
avg	**64.69**	**20.45**	-	**21.69**	-	**54.04**	-

TABLE 10. Accuracy comparison of cross-dataset testing for FNN models, ordered by dataset id, expressed in percentages.

id	Original EM	Original EM	Original	Shared EM	Shared Delta	Quad EM	Quad Delta
	accuracy	cross-testing	Delta	cross-testing		cross-testing	
1	23.69	73.20	49.51	73.37	49.68	85.17	61.48
2	73.93	73.17	-0.76	73.07	-0.86	84.23	10.31
3	84.28	50.17	-34.11	51.87	-32.41	67.93	-16.35
4	48.17	73.33	25.16	73.30	25.13	84.23	36.06
5	28.65	88.03	**59.38**	86.60	**57.95**	94.40	**65.75**
6	50.40	73.50	23.10	73.33	22.93	84.37	33.97
7	56.94	73.50	16.56	73.33	16.40	84.33	27.40
8	72.35	73.33	0.98	73.17	0.82	84.37	12.02
avg	**54.8**	**72.28**	-	**72.25**	-	**83.63**	-

TABLE 11. Accuracy comparison of cross-dataset testing for BERT models, ordered by dataset id, expressed in percentages.

id	Original EM	Original EM	Original	Shared EM	Shared Delta	Quad EM	Quad Delta
	accuracy	cross-testing	Delta	cross-testing		cross-testing	
1	37.41	80.63	43.22	81.10	43.69	90.87	53.45
2	93.25	78.93	-14.32	80.07	-13.18	90.13	-3.12
3	94.17	56.87	-37.31	58.27	-35.91	76.37	-17.81
4	71.06	79.30	8.24	80.33	9.28	90.27	19.21
5	41.76	91.93	**50.17**	92.23	**50.47**	97.90	**56.14**
6	43.20	79.47	36.27	80.43	37.23	90.40	47.20
7	93.79	79.30	-14.49	80.37	-13.42	90.33	-3.46
8	88.89	79.23	-9.66	80.30	-8.59	90.33	1.44
avg	**70.44**	**78.21**	-	**79.14**	-	**89.58**	-

While label reduction can yield valuable insights into model behavior and enhance computational efficiency, researchers must remain cognizant of the potential trade-offs, such as reduced model expressiveness and generalization capabilities.

Research in TED aims to develop models capable of consistently predicting emotions in diverse text samples, thereby demonstrating consistent performance across unseen data encountered "in the wild." Cross-dataset testing is a crucial evaluation metric for assessing the model's performance under real-world conditions. In this context, the subsequent phase of the experiments focused on cross-dataset testing as the authentic benchmark of model performance.

A. CROSS-DATASET TESTING

The performance of trained models for predicting emotions in unseen data varies considerably. When tested on different datasets, the models showed significant changes in performance (over 5% in either increase or decrease) as depicted in the delta columns from Tables 9, 10, and 11. These tables illustrate the changes in performance resulting from cross-dataset testing with different emotion mappings. Delta columns show the differences in performance. The first delta column quantifies the change in accuracy when trained models were presented with unseen data from other datasets without any emotion mapping equivalences. The second delta column represents the change in accuracy when the models used for cross-dataset testing used the shared emotions mapping. Finally, the third delta column shows the changes in accuracy when models trained with the quadrant emotion mappings were used during cross-dataset testing.

1) CROSS-DATASET TESTING IN LINEAR REGRESSION MODELS

The observed variations in cross-dataset testing performance underscore potential limitations arising from using linear regression models in the context of NLP and TED, with a particular sensitivity to the number of classifiers (labels in the EM). Linear regression's core assumption of a linear relationship between features and target variables may become increasingly strained as the number of emotion labels varies. More distinct emotion categories introduce a greater need to model complex, potentially non-linear boundaries between emotions within the feature space.

The subjective labeling of emotional content and linguistic variation between different text corpora further exacerbates this effect. Datasets relying on fine-grained emotion labels may have subtle linguistic cues and patterns distinguishing these emotions, which a linear model struggles to represent adequately. Similarly, when a model has learned decision boundaries based on a specific set of labels, it can be ill-equipped to deal with a differing label set found in cross-dataset testing.

For the present study, results for the cross-dataset testing in the linear regression models (table 9) show an average of 20.45% accuracy while using the original EM, 21.69% with the shared emotions mapping and 54.04% for the quadrant emotions. These measurements contrast with the original dataset's accuracy average of 64.69%

The most significant performance change is found on dataset id 7 using its original EM, demonstrating an 80.4% decrease in accuracy when evaluating other datasets. When using the shared emotions mapping, the most significant change is from the model trained with dataset id 3, with a 73.18% decrease in accuracy. Finally, the most considerable variation for the quadrant emotions mapping is present in the model trained with dataset id 8, with a decrease in accuracy of 36.78%. 23 of the 24 cross-dataset testing measurements present a significant (above 5%) change in performance for the models with linear regression algorithms. The only non-significant difference (less than 5%) was reported by the model trained with dataset id 6 with the quadrant emotions mapping, with a variation of only a 1.47% increase.

2) CROSS-DATASET TESTING IN FNN MODELS

In contrast to the findings observed with linear regression models, using FNNs revealed some cases of a notable increase in accuracy during cross-dataset testing with a reduction in emotion labels. This outcome underscores the impact of label manipulation and the inherent characteristics of NNs on model performance.

FNNs inherently excel in modeling non-linear relationships, a crucial aspect for effectively capturing subtle linguistic nuances associated with emotions. Their layered architecture facilitates intricate feature transformations, potentially making them less susceptible to dataset-specific variations than linear models. Moreover, implicit regularization mechanisms during training can help mitigate overfitting concerns inherent in smaller datasets.

The significant enhancement in cross-dataset accuracy observed with label reduction suggests potential inconsistencies or semantic discrepancies across the original sets of emotion labels. By leveraging their enhanced representational capabilities, NNs may identify shared emotional concepts despite diverse labeling schemes. Furthermore, aligning label reduction with established psychological constructs, such as optimized emotion categories (and, in this way, tuning the used EM), can encourage models to learn fundamental emotion signals that are less susceptible to variations across datasets.

The improvement in cross-dataset testing outcomes suggests the existence of shared underlying representations of emotion within the datasets. It is plausible that the complexity of multiple labels led models to overfit specific dataset characteristics, while label reduction inadvertently emphasized transferable patterns of emotional expression. Ultimately, NNs appear adept at identifying core linguistic markers of emotion regardless of nuanced labeling distinctions.

In other words, these results suggest that having more labels does not necessarily mean that the dataset is of better quality, and therefore, the trained model accuracy might be compromised.

For the FNN models (table 10), the most significant change was presented by dataset 5, comparing its performance using its own EM against the rest of the unseen datasets, with a 59.38% increase in accuracy.

Out of the 24 different cross-dataset testing scenarios, 4 reported non-significant changes in accuracy (less than 5%), while the remaining 20 did show changes above 5%.

3) CROSS-DATASET TESTING IN BERT-BASED MODELS

Integrating BERT-based classifiers in the experiments exhibited significant enhancements in baseline accuracy, further augmented by notable advancements during cross-dataset testing. This outcome, particularly when coupled with label reduction, underscores the intricate relationship between BERT's attributes and the manipulation of emotion labels.

BERT models derive their efficacy from ample pre-training on extensive text corpora utilizing masked language modeling and next-sentence prediction methodologies. This fosters a profound contextual comprehension of language, resulting in the generation of nuanced word representations. It is plausible that these representations empower a BERT classifier to grasp the underlying semantic connections among diverse emotion labels across datasets, even in the presence of misaligned labeling schemes.

Moreover, BERT's attention mechanisms facilitate sensitivity to context-specific word interactions, aligning well with the expression of emotions through intricate linguistic combinations, thereby potentially bridging the disparities introduced by disparate emotion labels.

It is imperative to underscore that adjusting the number of labels should not be considered a trivial modification. These experiments substantially impact performance while upholding consistent datasets, preprocessing techniques, and algorithmic selections. Reducing label sets may compel BERT models to attend to broader contextual emotional expression patterns that rely less on dataset-specific labeling peculiarities. However, excessive abstraction in label reduction poses the risk of forfeiting discriminative information, thereby diminishing the capacity to discern pertinent nuances of emotion. Conversely, employing an excessive number of labels, particularly those that are overly specific or granular, can introduce variability and noise during model training. This can lead to difficulties for BERT classifiers in identifying meaningful patterns and associations amidst many fine-grained emotional distinctions. As a result, the model may struggle to generalize effectively to unseen data, thereby reducing overall accuracy and robustness.

The augmentation in cross-dataset accuracy implies that BERT's robust linguistic pre-training and contextual emphasis enable it to circumvent certain limitations associated with strict overfitting to individual dataset characteristics. These findings underscore the pivotal role of nuanced semantic comprehension in navigating the intrinsic subjectivity and variability of emotion labeling in textual data.

For the BERT model (table 11) dataset 5 and its comparison using its own EM against unseen data is also the biggest change in performance, with an increase of 50.17%. Out of the 24 accuracy results, only three of them showed changes of less than 5% variation. The remaining 21 results showed significant, above 5% variations.

During all the cross-dataset testings, we found that out of 72 experiments (24 for each of linear regression, FNN, and BERT), in 8 cases, the variance in accuracy was non-significant (less than 5%), while in the remaining 64 cases, the change was significant. This means that for around 88% of the experiments with all elements equal, the introduced emotion mappings significantly impacted performance and generalizability.

V. CONCLUSION

For the research question: How does the arbitrary selection of an emotion model affect the performance and generalizability of machine learning models for text emotion detection, and how can emotion labels from different models be standardized to improve compatibility and generalization across diverse datasets? This investigation provides compelling evidence of the significant influence exerted by the arbitrary selection of EMs on the effectiveness and adaptability of ML models employed in TED. When all other experimental variables remained constant, alterations in emotion mappings consistently induced model accuracy and performance fluctuations across different datasets. Notably, in 88% of the conducted experiments, these fluctuations surpassed the predetermined threshold of 5%, thus substantiating the rejection of the null hypothesis.

We propose that integrating standardized emotion labels as a hyper-parameter for TED research is pivotal in fostering compatibility across datasets and enhancing overall model generalization. Models trained with standardized labels consistently performed better than those trained with dataset-specific tags. These outcomes underscore the critical importance of deliberate and informed decisions regarding the choice of EMs and the implementation of standardized labeling practices in developing robust TED systems that would perform consistently in both controlled environments and "on the wild" with unseen text.

The implications of these findings extend beyond academic circles, impacting developers and practitioners working on affective-aware applications. As the demand for sophisticated text-based emotional analysis grows (be it in customer service bots, mental health assessments, or personalized content delivery), so does the necessity for robust and adaptable TED systems. This research provides a clear directive for future developments by integrating standardized emotion labels and fostering a more systematic approach to EM selection. This, in turn, paves the way for more reliable and nuanced emotional analysis tools, which are crucial for advancing human-computer interaction in increasingly digital landscapes.

In summary, this research delivers valuable insights for practitioners seeking to enhance the efficacy and adaptability of ML models operating within emotionally nuanced text datasets. Furthermore, it prompts pertinent inquiries regarding the necessity for unified theoretical frameworks governing the modeling and labeling of emotions within the computational linguistics domain.

VI. SCOPE AND LIMITATIONS

The experimentation encompassed three distinct models: linear regression, FNN, and BERT. While the datasets utilized were publicly accessible, it is essential to note that they do not represent an exhaustive compilation. The source code implementation of all algorithms has been disclosed along with this paper; however, it is conceivable that practitioners may discover alternative, more optimized approaches. Nonetheless, it is essential to clarify that the primary objective of this research was not to identify an optimal implementation nor to delve into aspects such as text preprocessing, model architecture, programming methodology, or other ancillary factors. Instead, the focus remained on demonstrating the impact of alterations in the emotion model (EM) while maintaining all other variables constant and adhering to a straightforward methodology.

All datasets employed in the study were exclusively in English, as were the corresponding labels. The experimentation used Google Colab Pro, leveraging prioritized access to high-performance NVIDIA A100 GPUs. While the specific specifications of the allocated compute instance may vary within Google's cloud environment, it is noteworthy that A100 GPUs consistently offered substantial acceleration and memory capacity conducive to model training and evaluation. Nevertheless, it is pertinent to acknowledge that variations in computational resources may influence the reported findings, as previously stated in this document.

REFERENCES

[1] I Gartner. (2019). *2021 Strategic Roadmap for Enterprise AI: Natural Language Architecture*. [Online]. Available: https://www.gartner.com/en/documents/3994504

[2] H. Li, B. X. B. Yu, G. Li, and H. Gao, "Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews," *Tourism Manage.*, vol. 96, Jun. 2023, Art. no. 104707.

[3] I Gartner. (2021). *Hype Cycle for Natural Language Technologies, 2021*. [Online]. Available: https://www.gartner.com/en/documents/3994504

[4] P. Vyas, M. Reisslein, B. P. Rimal, G. Vyas, G. P. Basyal, and P. Muzumdar, ''Automated classification of societal sentiments on Twitter with machine learning,'' *IEEE Trans. Technol. Soc.*, vol. 3, no. 2, pp. 100–110, Jun. 2022.

[5] P. Nandwani and R. Verma, ''A review on sentiment analysis and emotion detection from text,'' *Social Netw. Anal. Mining*, vol. 11, no. 1, p. 81, Aug. 2021.

[6] P. S. Sreeja and G. Mahalakshmi, ''Emotion models: A review,'' *Int. J. Control Theory Appl.*, vol. 10, no. 8, pp. 651–657, 2017.

[7] R. Plutchik, *Emotion: A Psychoevolutionary Synthesis*. Manhattan, NY, USA: Harper & Row, 1980.

[8] C. E. Izard, *Human Emotions*. New York, NY, USA: Springer, 1977.

[9] P. Ekman, ''An argument for basic emotions,'' *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, 1992.

[10] J. A. Russell, ''A circumplex model of affect,'' *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, 1980.

[11] K. R. Scherer, ''Emotions: Definition, nature, and general issues,'' in *Handbook of Emotions*. New York, NY, USA: Guilford Press, 2000, pp. 137–154.

[12] J. J. Gross, ''The component process model of emotion regulation,'' *Psychol. Inquiry*, vol. 9, no. 2, pp. 80–99, 1998.

[13] A. Mehrabian, ''Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,'' in *Current Psychology*. New York, NY, USA: Springer, 1996.

[14] T. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[15] D. Maulud and A. M. Abdulazeez, ''A review on linear regression comprehensive in machine learning,'' *J. Appl. Sci. Technol. Trends*, vol. 1, no. 44, pp. 140–147, Dec. 2020.

[16] B. Duncan and Y. Zhang, ''Neural networks for sentiment analysis on Twitter,'' in *Proc. IEEE 14th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI*CC)*, Jul. 2015, pp. 275–278. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7259397

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol. 1, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[18] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, ''Automatic differentiation in Pytorch,'' in *Proc. NIPS-W*, 2017, pp. 1–4.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, ''Scikit-learn: Machine learning in Python,'' *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jul. 2011.

[20] S. M. Basha and D. Rajput, *Survey on Evaluating the Performance of Machine Learning Algorithms: Past Contributions and Future Roadmap*. Cambridge, MA, USA: Academic, pp. 153–164, Jan. 2019.

[21] A. Alslaity and R. Orji, ''Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions,'' *Behaviour Inf. Technol.*, vol. 43, pp. 1–26, Dec. 2022.

[22] J. Herzig, M. Shmueli-Scheuer, and D. Konopnicki, *Emotion Detection From Text via Ensemble Classification Using Word Embeddings*. New York, NY, USA: ACM, Oct. 2017.

[23] W. Witon, P. Colombo, A. Modi, and M. Kapadia, ''Disney at IEST 2018: Predicting emotions using an ensemble,'' in *Proc. 9th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, A. Balahur, S. M. Mohammad, V. Hoste, and R. Klinger, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 248–253. [Online]. Available: https://aclanthology.org/W18-6236

[24] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, and H. Prendinger, ''Decision support with text-based emotion recognition: Deep learning for affective computing,'' *Decis. Support Syst.*, vol. 115, pp. 24–35, Mar. 2018.

[25] F. Anzum and M. L. Gavrilova, ''Emotion detection from microblogs using novel input representation,'' *IEEE Access*, vol. 11, pp. 19512–19522, 2023.

[26] P. Probst, A.-L. Boulesteix, and B. Bischl, ''Tunability: Importance of hyperparameters of machine learning algorithms,'' *J. Mach. Learn. Res.*, vol. 20, no. 1, Jan. 2019, Art. no. 19341965.

[27] P. Lin, X. Luo, and Y. Fan, ''A survey of sentiment analysis based on deep learning,'' *Int. J. Comput. Inf. Eng.*, vol. 14, no. 12, pp. 473–485, 2020.

[28] A. D. L. Languré and M. Zareei, ''Breaking barriers in sentiment analysis and text emotion detection: Toward a unified assessment framework,'' *IEEE Access*, vol. 11, pp. 125698–125715, 2023.

[29] C. Cortes, L. D. Jackel, and W.-P. Chiang, ''Limits on learning machine accuracy imposed by data quality,'' in *Proc. Knowl. Discovery Data Mining*, 1995, pp. 1–8.

[30] M. Stone, ''Cross-validatory choice and assessment of statistical predictions,'' *J. Roy. Stat. Soc., Ser. B Methodol.*, vol. 36, no. 2, pp. 111–133, 1974.

[31] G. M. Shafiq, T. Hamza, M. F. Alrahmawy, and R. El-Deeb, ''Enhancing Arabic aspect-based sentiment analysis using end-to-end model,'' *IEEE Access*, vol. 11, pp. 142062–142076, 2023.

[32] CrowdFlower. (2016). *Emotion Detection From Text*. [Online]. Available: https://data.world/crowdflower/sentiment-analysis-in-text

[33] P. Govindaraj. (2020). *Emotions Dataset for NLP*. Kaggle. [Online]. Available: https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp/data

[34] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, ''CARER: Contextualized affect representations for emotion recognition,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.* Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 3687–3697. [Online]. Available: https://www.aclweb.org/anthology/D18-1404

[35] E. S. Dan-Glauser and K. R. Scherer, ''The difficulties in emotion regulation scale (DERS): Factor structure and consistency of a French translation,'' *Swiss J. Psychol.*, vol. 72, no. 1, pp. 5–11, Jan. 2013, doi: 10.1024/1421-0185/a000093.

[36] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, ''GoEmotions: A dataset of fine-grained emotions,'' in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 4040–4054.

[37] C. Strapparava and R. Mihalcea, ''Semeval-2007 task 14: Affective text,'' in *Proc. 4th Int. Workshop Semantic Evaluations (SemEval)*, 2007, pp. 70–74.

[38] D. Ghazi, D. Inkpen, and S. Szpakowicz, ''Detecting emotion stimuli in emotion-bearing sentences,'' in *Proc. 16th Int. Conf.*, Cairo, Egypt. Cham, Switzerland: Springer, 2015, pp. 152–165.

[39] S. M. Mohammad and F. Bravo-Marquez, ''WASSA-2017 shared task on emotion intensity,'' in *Proc. Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal. (WASSA)*, Copenhagen, Denmark, 2017, pp. 65–77.

ALEJANDRO DE LEÓN LANGURÉ received the bachelor's degree in computer science from the School of Engineering and Sciences, Tecnológico de Monterrey, and the master's degree from the School of Economics and Business, Anahuac University. He is currently pursuing the Ph.D. degree in computer sciences with the School of Engineering and Sciences, Tecnológico de Monterrey, with a focus on machine learning for affective computing and text emotion detection.

MAHDI ZAREEI (Senior Member, IEEE) received the M.Sc. degree in computer networks from the University of Science Malaysia, in 2011, and the Ph.D. degree from the Communication Systems and Networks Research Group, Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Malaysia, in 2016. In 2017, he joined the School of Engineering and Sciences, Tecnológico de Monterrey, as a Postdoctoral Fellow, where he was a Research Professor, in 2019. His research interests include wireless sensor and ad hoc networks, energy harvesting sensors, information security, and machine learning. He is a member of Mexican National Researchers System (Level I). He is an Associate Editor of IEEE Access and *PLOS One*.