IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 13, 2021, accepted January 5, 2022, date of publication January 14, 2022, date of current version January 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3143819

Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods

IQRA AMEER[1], GRIGORI SIDOROV[1], HELENA GÓMEZ-ADORNO[2], AND RAO MUHAMMAD ADEEL NAWAB[3]

[1]Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Ciudad de México 07738, Mexico
[2]Instituto de Investigación en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México (UNAM), Ciudad de México 04510, Mexico
[3]Computer Science Department, COMSATS University Islamabad, Lahore Campus, Lahore 54000, Pakistan

Corresponding author: Grigori Sidorov (sidorov@cic.ipn.mx)

This work was supported in part by the Mexican Government of the CONACYT, Mexico, under Grant A1-S-4785; and in part by the Secretaría de Investigación y Posgrado, Instituto Politécnico Nacional, Mexico, under Grant 20211784, Grant 20211884, and Grant 20211178.

ABSTRACT The multi-label emotion classification task aims to identify all possible emotions in a written text that best represent the author's mental state. In recent years, multi-label emotion classification attracted the attention of researchers due to its potential applications in e-learning, health care, marketing, etc. There is a need for standard benchmark corpora to develop and evaluate multi-label emotion classification methods. The majority of benchmark corpora were developed for the English language (monolingual corpora) using tweets. However, the multi-label emotion classification problem is not explored for code-mixed text, for example, English and Roman Urdu, although the code-mixed text is widely used in Facebook posts/comments, tweets, SMS messages, particularly by the South Asian community. For filling this gap, this study presents a large benchmark corpus for the multi-label emotion classification task, which comprises 11,914 code-mixed (English and Roman Urdu) SMS messages. Each code-mixed (English and Roman Urdu) SMS message manually annotated using a set of 12 emotions, including anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral (no emotion). As a secondary contribution, we applied and compared state-of-the-art classical machine learning (content-based methods – three word n-gram features and eight character n-gram features), deep learning (CNN, RNN, Bi-RNN, GRU, Bi-GRU, LSTM, and Bi-LSTM), and transfer learning-based methods (BERT and XLNet) on our proposed corpus. After our extensive experimentation, the best results were obtained using state-of-the-art classical machine learning methods on word uni-gram (Micro Precision = 0.67, Micro Recall = 0.54, Micro F_1 = 0.67) with a combination of OVR multi-label and SVC single-label machine learning algorithms. Our proposed corpus is free and publicly available for research purposes to foster research in an under-resourced language (Roman Urdu).

INDEX TERMS Code-mixed text, Roman Urdu and English, deep learning, classical machine learning, multi-label emotion classification, SMS messages, transfer learning.

I. INTRODUCTION

A single piece of text may contain one or more emotions. A single-label emotion classification task aims to predict only one emotion of a text. The main drawback of single-label emotion classification is that it only captures one emotion in a given text, making it difficult to completely understand the author's emotional state. Multi-label emotion classification overcomes this limitation by capturing all possible emotions in a given text, see examples in Table 1. Consequently, we can make a more accurate judgment about the emotional state of an author. Multi-label emotion classification has potential applications in various domains. For example, in E-learning, multi-label emotion classification can adjust the learning techniques in conformity with the learner. Multi-label emotion classification can be helpful in health care to determine the feelings and comfort level of the patient towards the treatment. Multi-label emotion classification can be used, for example, in stock market monitoring or prioritizing calls in a call center.

The associate editor coordinating the review of this manuscript and approving it for publication was Khin Wee Lai.

TABLE 1. Examples of SMS messages with multi-label emotions from our proposed CM-MEC-21 corpus.

ID	SMS Message	English Translation	Emotions
1	Yaro phr Huda and Mara ki birthdays ka kya plan hai? I am excited :D	Dude, What is the plan of Huda & Mara's birthday then? I am Excited :D	Joy
2	Since the 'update' my iPhone loses power nearly 40% faster. Ye ho kia raha hai???	Since the 'update' my iPhone loses power nearly 40% faster. What is happening???	Anger, Disgust
3	Main Hp pavilion 15-n200 notebook pc series use kr raha hon	I am using Hp pavilion 15-n200 notebook pc series	Neutral
4	Please jaldi aa jao we are missing you	Please come soon we are missing you	Love, Joy, Trust

In general, code-mixing can be characterized as the use of two or more languages at the same time. According to [1], more than 50% of Europeans use other language besides their mother language. The internet is the most prominent source in promoting global, linguistic code-mixed culture. In South Asian community and particularly in Pakistan, code-mixed (English and Roman Urdu) text became a preferable script for Facebook comments/posts [2], tweets [3]–[5], and daily communication using SMS message [6]. It can be noted from these studies that the use of code-mixed digital text is increasing. Thus, there is a need to develop standard evaluation resources and methods for code-mixed texts for various applications, such as author profiling, sentiment analysis, emotion analysis, etc.

Standard evaluation resources are needed to develop, evaluate, and compare multi-label emotion classification methods. Previous studies developed few corpora for multi-label emotion classification using English tweets (monolingual) [7]–[10]. However, the problem of multi-label emotion classification is not explored for code-mixed (say, English and Roman Urdu) texts. To fulfill this research gap, the present study aims to develop a large benchmark code-mixed (English and Roman Urdu) corpus for the multi-label emotion classification task and evaluate it.

The two main objectives of this study are: (1) to develop a large benchmark code-mixed (English and Roman Urdu) SMS messages corpus for multi-label emotion classification task and (2) to apply, evaluate, and compare state-of-the-art classical machine learning, deep learning, and transfer learning methods on the proposed corpus to investigate most suitable methods for multi-label emotion classification on a code-mixed corpus. For the first objective, we developed a large benchmark multi-label emotion classification corpus, which contains 11,914 code-mixed (English and Roman Urdu) multi-label SMS messages, hereafter called CM-MEC-21 corpus. In the CM-MEC-21 corpus, each code-mixed SMS message is manually annotated from a predefined set of 12 emotions, which are: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral (no emotion). For the second objective, we developed and applied state-of-the-art classical machine learning, deep learning, and transfer learning methods on our proposed CM-MEC-21 corpus.

We believe that our proposed CM-MEC-21 corpus will be helpful for: (1) promotion of research in an under-resourced language, i.e., Roman Urdu, (2) development of bi-lingual dictionaries for English and Roman Urdu languages, (2) carrying out a detailed comparison of existing methods for the multi-label emotion classification task, and (4) development and evaluation of the new methods for multi-label emotion classification on code-mixed text (in our case, English and Roman Urdu).

The rest of this paper is organized as follows: Section II describes the existing multi-label emotion classification corpora and methods. Section III presents the corpus compilation process used to create the proposed corpus. Section IV describes methods for multi-label emotion classification task. Section V presents the experimental setup (dataset, techniques, evaluation methodology, and evaluation measures). Results and their analysis are presented in Section VI. Finally, Section VII concludes the paper and discusses potential avenues for future work.

II. RELATED WORK

In literature, efforts were made to develop benchmark corpora for the emotion classification task. However, the majority of these efforts focused on the single-label emotion classification task. One of the most prominent efforts of the single-label emotion classification task is a series of international competitions organized by SemEval [7]. The main outcome of these competitions is a collection of benchmark corpora, which can be used for the development, comparison, and evaluation of single-label classification methods. Other researchers also made efforts to develop corpora for single-label emotion classification task [11]–[19].

Regarding the multi-label emotion classification task, we found that only three benchmark corpora were developed. [7] developed a large benchmark monolingual (only one language) corpora of English, Spanish, and Arabic for SemEval-2018 multi-label emotion classification competition. The English corpus consists of 10,983 tweets (training instances = 6,838, validation instances = 886, and testing instances = 3,259) manually annotated by seven annotators with the presence/absence of 12 emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and neutral (no emotion). The best system [20] applied a multi-layer self-attention with the bidirectional long short-term memory architecture approach and achieved an accuracy of 58.80%.

In another effort towards the development of the benchmark corpus for the multi-label emotion classification task, [8] created the CBET corpus, which contains 81,162 English tweets manually annotated with nine emotions, including anger, fear, disgust, joy, love, sadness, surprise,

thankfulness, and guilt. The authors applied lexical and learning-based methods for multi-label emotion classification on the CBET corpus, and they achieved the best results on word uni-grams (Precision = 47.07).

A similar corpus is BMET corpus [9], which consists of 96,323 English tweets manually annotated for the multi-label emotion classification task using six emotions: anger, fear, joy, sadness, surprise, and thankfulness. The authors applied the Latent Variable Chain (LVC) Transformation model and achieved the best micro average F_1 score of 67.19%.

As can be observed from the multi-label emotion classification corpora mentioned above, all existing multi-label emotion classification corpora are developed for monolingual setup using tweets. Although SMS messages and tweets look similar, there are significant differences between them [6]. The first main difference between SMS messages and tweets is that usage of SMS messages is considerably wider spread than the usage of tweets. Out of the 8.97B worldwide connections of mobile, 5.67B connections are basic regular phones (non-smartphone) [21]. So, while contacting through texting, pure SMS messaging is the only possible option for individuals away from the Internet. Recently, organizations like WhatsApp, Facebook, Banks, etc., are using SMS message services to verify user profiles, security, and confirmation [6]. The second main difference between SMS messages and tweets is that most tweets are public, and SMS messages are private. As compared to the SMS communication medium [6], when a user broadcasts a tweet message on Twitter, Twitter provides open access to tweet settings, and anyone with an Internet connection can access and interact with them, or there are also followers-specific settings.

On the other hand, SMS messages are focused on private messaging such as a small audience, a one-to-one transmission, or group communication. Therefore, SMS messages' protocol can be called a personal, conversational medium, while tweets are more social. Researchers tend to pursue variations in texting patterns for public and personal circles. Considering the significant differences between SMS messages and tweets, we can conclude that tweets cannot be treated as an alternative to the SMS corpus, even though the tweets corpus is easier to compile.

To summarize, existing multi-label emotion classification corpora are mainly developed for the monolingual data (English) for tweets. However, there is no benchmark multi-label emotion classification corpus for code-mixed text, for example, English and Roman Urdu. This study presents a large code-mixed (English and Roman Urdu) corpus for the multi-label emotion classification task comprising 11,914 code-mixed SMS messages, manually annotated with 12 emotion categories. To the best of our knowledge, no such corpus was developed in the past.

III. CORPUS COMPILATION PROCESS

The primary objective of this research is to develop a large standard benchmark code-mixed (English and Roman Urdu) corpus for the multi-label emotion classification task. This section describes the source data, annotation process, characteristics, and standardization of our proposed corpus.

A. SOURCE DATA

To develop a large benchmark code-mixed corpus for the multi-label emotion classification task, we manually selected the data from an existing benchmark SMS-AP-18 corpus [6]. The SMS-AP-18 corpus was developed for code-mixed (English and Roman Urdu) author profiling based on SMS messages. Each document belongs to a profile of an author containing his collection of SMS messages. The SMS-AP-18 corpus contains 810 author profiles (a total of 84,694 code-mixed SMS messages). On average, there are 104.56 code-mixed SMS messages per profile. The SMS-AP-18 corpus is annotated with seven different author traits, including gender, age group, native city, native language, personality type, education level, and profession. The reason for selecting the SMS-AP-18 corpus for this study is that this is the only benchmark and publicly available corpus containing code-mixed (English and Roman Urdu) SMS messages written by individuals.

To develop our proposed CM-MEC-21 corpus, we manually selected a subset of 12,000 SMS messages from the first 168 author profiles of the SMS-AP-18 corpus. The reason behind selecting a subset of 12,000 code-mixed SMS messages is that it will be very challenging and time-consuming to annotate the entire SMS-AP-18 corpus for multi-label emotion classification on code-mixed SMS messages (English and Roman Urdu). We only selected those messages whose length is greater than or equal to five words. We did not perform any pre-processing on 12,000 code-mixed SMS messages because the SMS-AP-18 corpus was already pre-processed.

B. ANNOTATION PROCESS

This section presents the annotation process used to develop our proposed CM-MEC-21 corpus, including preparing annotation guidelines, annotations and calculating the Inter-Annotator Agreement (IAA).

1) ANNOTATION GUIDELINES

The proposed CM-MEC-21 corpus was manually annotated by three annotators (A, B, and C). All the annotators were graduates, experienced in-text annotations, and native Urdu speakers with high English language proficiency. To facilitate the annotation process of our proposed CM-MEC-21 corpus, we prepared a set of annotation guidelines. These annotation guidelines were used to manually annotate the code-mixed (English and Roman Urdu) SMS messages for multi-label emotion classification. Since this research focuses on multi-label emotion classification, annotators were asked to assign multiple labels (or emotions) to a code-mixed SMS message. However, the first label should be the most dominating among all the labels assigned to a code-mixed SMS message. If a code-mixed SMS message does not contain any emotion, only one label will be assigned, i.e.,

neutral. Following the set of emotion categories used in the SemEvel-2018 international competition on multi-label emotion classification on English tweets, annotators assigned one or more emotions to a code-mixed SMS message from the following twelve categories. The annotation process was performed using Excel files and completed in three months.

Definitions of the emotions are as follows:

- **Anger:** The emotion anger, also known as annoyance or rage, is an extreme emotional condition. It includes an awkward and bitter response to an anticipated incitement, danger, or hurt. [22]
- **Anticipation:** Anticipation is an emotion, including delight, excitement, or nervousness/anxiety because of or expecting an event [23], which also includes interest, hope, and prospect.
- **Disgust:** Disgust is a reaction to denial or refusal to something conceivably contagious [24], which also includes disinterest, loathing, and dislike.
- **Fear:** Fear is a feeling persuaded by an anticipated troublesome situation, or danger [25], which also includes anxiety, panic, and horror.
- **Joy:** Joy is a sensation of great pleasure, and happiness [26], which also includes ecstasy, pride, and delight.
- **Love:** Love encloses a range of positive and strong emotional states, from the magnificent virtue or good habit, the sound interpersonal affection, and the simplest pleasure [22], also includes adoration and affection.
- **Optimism:** Optimism is a mental attitude contemplating a trust or hope that the reaction of some particular aim, or conclusion in general, will be positive, supportive, and desirable [27], also includes confidence, certainty, and hopefulness.
- **Pessimism:** In general, pessimists likely to focus on the negatives of life, a depressed or negative mindset, also includes distrust, cynicism, and no confidence [28].
- **Sadness:** Sadness is a feeling of loss, hardship, and anguish [29], also includes pensiveness.
- **Surprise:** Surprise is a mental state that a person might feel if something unanticipated occurs [30], which also includes amazement and distraction.
- **Trust:** Usually refers to a circumstance defined by the following aspects: One group (trustor) is ready to depend on the activities of another group (trustee); the situation is directed to the future [31], also includes confidence, belief, and faith.
- **Neutral:** There is no emotion(s) in a sentence.

2) ANNOTATIONS

Annotations were performed by three annotators (A, B, and C). All the annotators were native speakers of Urdu and had a high level of proficiency in the English language.

Annotations were carried out in two steps. In the first step, a subset of 200 code-mixed (Roman Urdu and English) SMS messages was annotated by annotators A, B, and C using the annotation guidelines. Annotators discussed the annotations of 200 code-mixed SMS messages and revised the annotation guidelines to improve the quality of annotations further. In the second step, revised annotation guidelines were used to annotate the remaining 11,800 code-mixed SMS messages.

To select the set of gold standard labels (emotions) for each code-mixed SMS message, we used the following guidelines. A code-mixed SMS message was annotated with only those labels, which were assigned by at least two annotators. If an SMS message did not have a single label that at least two annotators assigned, it was discarded. Out of 12,000 code-mixed SMS messages, 86 were discarded, and the final gold standard corpus contained 11,914 code-mixed SMS messages.

3) INTER-ANNOTATOR AGREEMENT

After annotations, Cohen's Kappa Coefficient and inter-rater agreement were computed. We computed inter-rater agreement as to the percentage of times each pair of annotators agree. We achieved a moderate level of Cohen's Kappa Coefficient score of 0.620 and the inter-rater agreement of 0.618. It can be noted that inter-rater agreement scores are lower than SemEval-2018 (inter-rater agreement = 83.38) shared task on English tweets (mono-lingual) [32]. According to previous studies, human annotators were agreed only approximately 70-80% of the time for binary or ternary classification schemes, and the more classes there are, the more challenging it is for annotators to agree [33]–[36]. Our CM-EMC-21 corpus consists of 12 classes, and this highlights that code-mixed multi-label emotion annotation is a complex task for humans to agree.

C. CORPUS CHARACTERISTICS AND STANDARDIZATION

The proposed CM-MEC-21 corpus contains 11,914 code-mixed SMS messages. The minimum and maximum length of code-mixed SMS messages in the CM-MEC-211 corpus is 5 and 99 words, respectively. The average length of a code-mixed SMS message is 12 words. The proposed corpus contains a total of 141,997 words and 15,660 word types (unique words).

Table 2 shows the percentage of code-mixed (English and Roman Urdu) SMS messages annotated with a given type of emotion. The statistics in these rows sum up to more than 100% because a single code-mixed (English and Roman Urdu) SMS message may be annotated with more than one label (emotion). It can be noted that neutral, anticipation, joy, trust, and optimism labels got a higher percentage. Pessimism, anger, and surprise are rare emotions. Table 3 indicates the number of labels assigned to code-mixed (English and Roman Urdu) SMS messages. We standardized the proposed CM-MEC-21 corpus in CSV format and made it publicly available for research purposes.

IV. METHODS FOR MULTI-LABEL EMOTION CLASSIFICATION

To demonstrate how our proposed CM-MEC-21 corpus can be used to develop, evaluate and compare methods for the multi-label emotion classification task, we applied and

TABLE 2. Percentage of code-mixed (English and Roman Urdu) SMS messages that were annotated with a given emotion.

Language	anger	anti.	disg.	fear	joy	love	optm.	pessi.	sadn.	surp.	trust	neutral
Eng + R-U	2.8	13.34	9.39	4.78	12.9	2.79	11.36	2.14	6.32	3	12.93	33.37

TABLE 3. Number of emotion labels assigned to code-mixed (English and Roman Urdu) SMS messages.

# of Labels Assigned	# of Messages
1	10,309
2	1,400
3	187
4+	18

compared three main types of popular and widely used supervised machine learning methods: (1) state-of-the-art classical machine learning methods (content-based methods), (2) state-of-the-art deep learning methods, and (3) state-of-the-art transfer learning methods. As far as we know, no previous study made such a detailed and thorough comparison of state-of-the-art methods for the multi-label emotion classification task on the code-mixed SMS messages. Below we describe these methods in detail.

A. CONTENT-BASED METHODS

Words and characters help create contextual content. Their sequence and structure can give important insights to classify texts. In earlier studies, [37] used the content-based approach for emotion classification on the responses of psychologists and non-psychologist students. Wang [38] and Ameer *et al*. [39] applied content-based methods for automatic emotion identification in texts.

1) N-GRAM FEATURES

The content-based methods for emotion classification tasks are based on n-grams taken from the code-mixed (English and Roman Urdu) SMS messages. The term n-gram (of characters or words) refers to a series of sequential tokens in a sentence, paragraph, and document. A group of n-grams can be generated by considering a series of tokens moving over a string, considering one token at a time. The series can be of length 1 (uni-grams), length 2 (bi-grams), length 3 (tri-grams), etc. N-grams are very widely used in natural language processing. As several examples, let us mention the following research works. Mohammad [18] used word uni-grams and word bi-grams for emotion classification on a corpus of newspaper headlines. Mohammad *et al*. [40] applied the word uni-grams and bi-grams along with punctuation marks, elongated words, emotion lexicons, and negation features to detect the emotional state and the stimulus of the authors of tweets on a corpus of 2012 US presidential elections. Mohammad *et al*. [13] used word n-grams, elongated words, and features associated with emotions for the emotion detection task. Content-based methods are also used in other tasks, for example, the author profiling task [6], [41]–[44].

Our study applied character and word n-grams for the multi-label emotion classification task on code-mixed SMS messages (English and Roman Urdu). We used TF-IDF values for n-grams, which is the most common solution (Scikit-learn was used for the implementation of the models). The maximum number of features for each experiment was 1,000, i.e., we used n-grams with the highest TF-IDF values. The length of *n-grams* was from 1 to 3 for word n-grams and from 3 to 10 for character n-grams, which are commonly used values.

B. DEEP LEARNING-BASED METHODS

The second type of methods is a deep learning-based methods, in which different state-of-the-art neural network models were applied [45], [46]. The series of SemEval [7], [20] emotion classification competitions played an important role in the development of emotion classification. We noticed that the most widely used models were Convolutional Neural Network, Recurrent Neural Networks, Long short-term memory, and Bidirectional Long short-term memory.

Rana [47] explored Gated Neural Networks for emotion classification from Noisy Speech and achieved promising results. Kim [48] trained a CNN model for text emotion classification and obtained a good classification effect. We applied seven state-of-the-art deep learning models for multi-label emotion classification on code-mixed (English and Roman Urdu) SMS messages, including Long short-term memory, Bidirectional Long short-term memory, Gated Neural Networks, Bidirectional Gated Neural Networks, Recurrent Neural Networks, Bidirectional Recurrent Neural Networks, and Convolutional Neural Network for multi-label emotion classification.

We used Scikit-learn implementation of deep learning models considering the following parameters, which are usually the default: hidden layers = 3, hidden units = 64, no. of epochs = 10, batch size = 64, and dropout = 0.001. The parameters of CNN model are as follows: activation function = Rectified Linear Units (ReLU), optimizer = adam, hidden layers = 3, loss function = sigmoid, no. of epochs = 10, batch size = 64, dropout = 0.001.

C. TRANSFER LEARNING-BASED METHODS

Bidirectional Encoder Representations from Transformers (BERT) [49] is one of the most popular advanced techniques for NLP problems. The BERT model provided state-of-the-art performance across various NLP tasks without any significant task-specific architecture alterations. BERT was primarily employed in aspect-based sentiment analysis, such as in [50]–[52]. Several other studies focused on emotion analysis using BERT. For example, in [53], the authors conducted a comparative analysis of multiple pre-trained transformer models for the text emotion recognition problem, including BERT. However, our study differentiates from the earlier one as we assess the emotion classification model's performance on multi-label code-mixed SMS messages, which is significantly more challenging.

In our study, the pre-trained uncased version of the BERT base model was applied, i.e., before the word tokenization step, the text is transformed to lowercase. The BERT Base model comprises 12 encoders, each with eight layers: four multi-head self-attention and four feed-forward layers.

The pre-trained XLNet was also applied in this work. The architecture of the XLNet base model consists of 12 transformer layers with 768 hidden layers, and 12 attention head layers were used. The XLNet tokenizer was used to split the sequences into tokens. After that, the tokens were padded, and classification was performed.

For the multi-label emotion classification task on our proposed CM-MEC-21 corpus, we added a fully connected layer and a sigmoid layer to these models. The batch size and learning rate were set to 32 and 2e-5, respectively. The models were optimized using the Adam optimizer, and the loss parameter was set to BCEWithLogitsLoss. The models were trained for ten epochs.

V. EXPERIMENTAL SETUP

This section describes how our proposed CM-MEC-21 corpus can be used to develop and evaluate emotion classification methods. The following sections present the dataset, techniques, evaluation methods, and evaluation measures in detail.

A. DATASET

Our proposed CM-MEC-21 corpus contains 11,914 SMS messages in the following three languages:

Code-Mixed (Roman Urdu and English): This Means the SMS contains text in both Roman Urdu and English languages. In the CM-MEC-21 corpus, there are 7, 474 code-mixed (Roman Urdu and English) SMS messages.

Roman Urdu: The SMS contains text in Roman Urdu language. In the CM-MEC-21 corpus, there are 3, 814 Roman Urdu SMS messages.

English: The SMS contains text in the English language. In the CM-MEC-21 corpus, there are 625 English SMS messages.

For experiments presented in this study, we divided our CM-MEC-21 corpus into three sub-corpora: (1) CM-CM-MEC-21 corpus, which contains 7, 474 code-mixed (Roman Urdu and English) SMS messages, (2) RM-CM-MEC-21 corpus, which contains 3, 814 SMS messages in Roman Urdu language, and (3) E-CM-MEC-21 corpus, which contains 625 SMS messages in the English language. Each sub-corpus is split into train-set and test-set with a ratio of 80%-20%. Out of 7, 474 instances in the CM-CM-MEC-21 sub-corpus, the train-set (called Train-CM-CM-MEC-21) contains 5, 980 instances and test-set (called Test-CM-CM-MEC-21) contains 1, 495 instances. From 3, 814 instances in the RU-CM-MEC-21 sub-corpus, the train-set (called Train-RU-CM-MEC-21) contains 3, 047 instances and test-set (called Test-RU-CM-MEC-21) contains 766 instances. Out of 625 instances in the E-CM-MEC-21 sub-corpus, the train-set (called Train-E-CM-MEC-21) contains 492 instances, and the test-set (called Test-E-CM-MEC-21) contains 133 instances.

B. TECHNIQUES

We applied classical machine learning, deep learning, and transfer (see Section IV) techniques on three sub-corpora. Below we describe the four sets of experiments (Exp1, Exp2, Exp3, and Exp4) that were designed and carried out with different combinations of training data and testing data present in the three sub-corpora.

Exp1. For this experiment, we used all three sub-corpora. For training, we combined Train-E-CM-MEC-21 and Train-RM-CM-MEC-21, whereas, for testing, we used the Test-CM-CM-MEC-21.

Exp2. For this experiment, we used one sub-corpus. For training, we used Train-CM-CM-MEC-21, and for testing, we used Test-CM-CM-MEC-21.

Exp3. For this experiment, we used all three sub-corpora. For training, we combined Train-E-CM-MEC-21 and Train-RM-CM-MEC-21, whereas, for testing, we combined Test-E-CM-MEC-21, Test-RM-CM-MEC-21, and Test-CM-CM-MEC-21.

Exp4. For this experiment, we used all three sub-corpora. For training, we combined Train-E-CM-MEC-21, Train-RM-CM-MEC-21, and Train-CM-CM-MEC-21, whereas, for testing, we combined Test-E-CM-MEC-21, Test-RM-CM-MEC-21, and Test-CM-CM-MEC-21.

The reason for designing Exp1, Exp2, Exp3, and Exp4 was to investigate the performance of different techniques on different types of training and testing datasets.

C. EVALUATION METHODOLOGY

The multi-label emotion classification problem on code-mixed SMS messages (English and Roman Urdu) is treated as a supervised multi-label text classification problem. We applied two different machine learning classification methods, including One vs. Rest and One vs. One, along with base classifiers including Random Forest, Logistic Regression, Naïve Bayes, Support Vector Machine, Bagging, and AdaBoost. The features extracted using content-based methods (see Section IV-A) are used as input in training and testing phases.

D. EVALUATION MEASURES

To evaluate the performance of multi-label emotion classification methods, three different evaluation measures were used: (i) Micro Precision, (ii) Micro Recall, (iii) Micro Avg. F_1.

Micro Precision is defined as precision of the aggregated contributions of all classes as (1), shown at the bottom of the next page, where:

E = set of 12 emotions,
e = emotion class.

Micro recall is defined as recall of the aggregated contributions of all classes as (2), shown at the bottom of the page.

Micro F_1 is the harmonic mean of Micro Precision (Mi_P) and Micro Recall (Mi_R):

$$\text{Micro } F_1 = \frac{2 \times Mi_P \times Mi_R}{Mi_P + Mi_R} \tag{3}$$

VI. RESULTS AND ANALYSIS

Tables 4, 5, 6, and 7 present the Micro Average Precision (Mi_P), Micro Average Recall (Mi_R) and Micro Average F_1 results obtained by applying state-of-the-art classical machine learning, deep learning, and transfer learning methods on our proposed CM-MEC-21 corpus. In these Tables, ''Features'' refers to the N-gram-based features of words and characters used for multi-label emotion classification. The ''ML Algorithms'' refers to a classical multi-label machine learning algorithm. ''RF'' refers to Random Forest ''NB'' refers to Naive Bayes, ''SVC'' refers to Support Vector Classifier. ''OVR'' refers to one-vs-rest and ''OVO'' one-vs-one multi-label machine learning algorithm. The ''DL Algorithms'' means deep learning algorithms used in this study such as Long short-term memory (LSTM), Bidirectional Long short-term memory (Bi-LSTM), Gated Neural Networks (GNN), Bidirectional Gated Neural Networks (Bi-GNN), Recurrent Neural Networks (RNN), Bidirectional Recurrent Neural Networks (Bi-RNN), and Convolutional Neural Network (CNN). The ''TL Algorithms'' refers to state-of-the-art transfer learning algorithms applied to evaluate our corpus, i.e., Bidirectional Encoder Representations from Transformers (BERT) and XLNet.

In Exp1 (see Table 4), overall, best results are obtained using word 2-grams (Micro Precision = 0.64, Micro Recall = 0.50, Micro F_1 = 0.64). The results show that word 2-grams are the most suitable features when we combine English (monolingual) and Roman Urdu (monolingual) SMS messages for training (Train-E-CM-MEC-21 + Train-RU-CM-MEC-21) and test them on the code-mixed SMS messages data (Test-CM-CM-MEC-21). It can be noted that the Micro F_1 score of 0.64 is not very high, highlighting the fact that multi-label emotion classification on code-mixed text is a challenging task. The best results obtained using deep learning (Micro F_1 = 0.30) and transfer learning methods (Micro F_1 = 0.26) are low. The possible reason for obtaining low results with deep learning and transfer learning is that the amount of data used for training is very small. Normally, deep learning and transfer learning methods require a huge amount of data for good training.

In Exp2 (see Table 5), overall, best results are obtained using character 3-grams (Micro Precision = 0.66, Micro Recall = 0.55, Micro F_1 = 0.66). This shows that character 3-grams are the most appropriate features when we train on code-mixed (English + Roman Urdu) SMS messages dataset (Train-CM-CM-MEC-21) and test them on code-mixed SMS messages dataset (Test-CM-CM-MEC-21). It can be observed that the Micro F_1 score of Exp2 (F_1 = 0.66) is slightly better than Exp1 (F_1 = 0.64). This highlights that the training model by combining data in two languages (Train-E-CM-MEC-21 + Train-RM-CM-MEC-21) and testing on code-mixed (English + Roman Urdu) data (Test-CM-MEC-21) did not have a significant effect on the performance. However, the best results were obtained with different features (word-2-grams in Exp1 and character-3-grams in Exp2). Similar to Exp1, results obtained with deep learning and transfer learning are low.

In Exp3 (see Table 6), overall, best results are obtained using word 1-gram (Micro Precision = 0.67, Micro Recall = 0.54, Micro F_1 = 0.67). The shows that word 1-grams are the most suitable features when we trained the model on code-mixed (English + Roman Urdu) SMS messages dataset (Train-CM-CM-MEC-21) and tested on a combination of English (monolingual), Roman Urdu (monolingual), and code-mixed SMS messages (Test-E-CM-MEC-21 + Test-RU-CM-MEC-21 + Test-CM-CM-MEC-21). Interestingly, these results are similar to the ones obtained in Exp2. This indicates that the model trained on code-mixed data performs equally well on combined English (monolingual), Roman Urdu (monolingual), and code-mixed (English + Roman Urdu) testing data. However, the best features are different in both experiments (char 3-grams in Exp 2 and word 1-grams in Exp 3). Again, similar to Exp1 and Exp2, results obtained using deep learning and transfer learning methods are low.

In Exp4 (see Table 7), overall, best results are obtained using word 1-grams (Micro Precision = 0.60, Micro Recall = 0.57, Micro F_1 = 0.60). This shows that word 1-grams are the most suitable features when we combine English (monolingual), Roman Urdu (monolingual), and code-mixed (English + Roman Urdu) SMS messages (Train-E-CM-MEC-21 + Train-RU-CM-MEC-21 + Train-CM-CM-MEC-21) and test them on combined English (monolingual), Roman Urdu (monolingual), and code-mixed (English + Roman Urdu) SMS messages dataset (Test-E-CM-MEC-21 + Test-RU-CM-MEC-21 + Test-CM-CM-MEC-21). It can be noted, the Micro F_1 score of this experiment is quite low compared to Exp 3 (Micro F_1 = 0.67), although both are tested on the same dataset. This indicates

$$Mi_P = \frac{\sum_{e \in E} \text{no. of messages correctly assigned to emotion class e}}{\sum_{e \in E} \text{no. of messages assigned to emotion class e}}, \tag{1}$$

$$Mi_R = \frac{\sum_{e \in E} \text{no. of messages correctly assigned to emotion class e}}{\sum_{e \in E} \text{no. of messages in emotion class e}} \tag{2}$$

TABLE 4. Results obtained by applying various classical machine learning, deep earning and transfer learning techniques (Exp1).

Features	ML Algorithms	Mi_P	Mi_R	Micro F_1
Classical Machine Learning				
Word 1-gram	OVR (Adaboost)	0.60	0.55	0.60
Word 2-grams	**OVR (Bagging)**	**0.64**	**0.50**	**0.64**
Word 3-grams	OVR (Bagging)	0.56	0.50	0.54
Char 3-grams	OVR (NB)	0.55	0.60	0.55
Char 4-grams	OVR (NB)	0.55	0.59	0.55
Char 5-grams	OVR (NB)	0.55	0.58	0.55
Char 6-grams	OVR (NB)	0.54	0.54	0.54
Char 7-grams	OVR (NB)	0.55	0.53	0.55
Char 8-grams	OVR (Bagging)	0.55	0.51	0.55
Char 9-grams	OVR (NB)	0.54	0.49	0.54
Char 10-grams	OVR (NB)	0.58	0.50	0.58
Deep Learning				
DL Algorithm		**Mi_P**	**Mi_R**	**Micro Avg. F_1**
LSTM		0.32	0.83	0.30
Bi-LSTM		0.32	0.82	0.30
GRU		0.32	0.83	0.30
Bi-GRU		0.41	0.73	0.28
RNN		0.31	0.82	0.30
Bi-RNN		0.32	0.82	0.30
CNN		0.34	0.61	0.27
Transfer Learning				
BERT		0.08	0.48	0.14
XLNet		0.44	0.18	0.26

*__Train Data__: Train-E-CM-MEC-21 (80%) + Train-RM-CM-MEC-21 (80%)

*__Test Data__: Test-CM-CM-MEC-21 (20%)

TABLE 5. Results obtained by applying various classical machine learning, deep earning and transfer learning techniques (Exp2).

Features	ML Algorithms	Mi_P	Mi_R	Micro F_1
Classical Machine Learning				
Word 1-gram	OVR (NB)	0.59	0.56	0.59
Word 2-grams	OVR (NB)	0.58	0.51	0.58
Word 3-grams	OVR (Bagging)	0.58	0.51	0.58
Char 3-grams	**OVR (SVC)**	**0.66**	**0.55**	**0.66**
Char 4-grams	OVR (SVC)	0.62	0.54	0.62
Char 5-grams	OVR (NB)	0.57	0.59	0.57
Char 6-grams	OVR (NB)	0.57	0.57	0.57
Char 7-grams	OVR (RF)	0.56	0.52	0.56
Char 8-grams	OVR (NB)	0.56	0.54	0.56
Char 9-grams	OVR (NB)	0.55	0.52	0.55
Char 10-grams	OVR (NB)	0.54	0.51	0.54
Deep Learning				
DL Algorithm		**Mi_P**	**Mi_R**	**Micro Avg. F_1**
LSTM		0.32	0.83	0.30
Bi-LSTM		0.39	0.81	0.27
GRU		0.32	0.75	0.30
Bi-GRU		0.36	0.76	0.26
RNN		0.31	0.82	0.30
Bi-RNN		0.39	0.81	0.27
CNN		0.34	0.66	0.25
Transfer Learning				
BERT		0.10	0.81	0.17
XLNet		0.46	0.29	0.35

*__Train Data__: Train-CM-CM-MEC-21 (80%)

*__Test Data__: Test-CM-CM-MEC-21 (20%)

that training models on code-mixed (Train-CM-MEC-21) data is more efficient compared to a combination of English (monolingual), Roman Urdu (monolingual), and code-mixed (English + Roman Urdu) SMS message on our proposed CM-MEC-21 corpus. Deep learning and transfer learning fail to produce promising results in the experiments.

Table 8 shows the best results obtained in Exp1, Exp2, Exp3, and Exp4. Overall, in all four experiments, content-based methods outperform the deep learning and transfer learning methods. This indicates that content-based methods were more efficient for multi-label emotion classification tasks when training data was small. It can be observed that results are not up to the mark for deep learning and transfer learning methods. This performance highlights that multi-label emotion classification on code-mixed (English and Roman Urdu) SMS messages is challenging for deep learning and transfer learning methods. Moreover, this performance indicates that complex deep learning and transfer learning models can find multi-label emotion classification tasks difficult when training data is small.

Regarding the length of n in content-based methods, it can be observed from summary Table 8 that the best performance

TABLE 6. Results obtained by applying various classical machine learning, deep earning and transfer learning techniques (Exp3).

Features	ML Algorithms	Mi_P	Mi_R	Micro F_1
Classical Machine Learning				
Word 1-gram	**OVR (SVC)**	**0.67**	**0.54**	**0.67**
Word 2-grams	OVR (SVC)	0.57	0.51	0.57
Word 3-grams	OVR (RF)	0.64	0.50	0.64
Char 3-grams	OVR (NB)	0.55	0.59	0.55
Char 4-grams	OVR (NB)	0.55	0.58	0.55
Char 5-grams	OVR (NB)	0.55	0.57	0.55
Char 6-grams	OVR (NB)	0.54	0.54	0.54
Char 7-grams	OVR (NB)	0.55	0.53	0.55
Char 8-grams	OVR (Bagging)	0.53	0.51	0.53
Char 9-grams	OVO (SVC)	0.57	0.51	0.57
Char 10-grams	OVR (NB)	0.56	0.50	0.56
Deep Learning				
DL Algorithm		**Mi_P**	**Mi_R**	**Micro Avg. F_1**
LSTM		0.33	0.81	0.31
Bi-LSTM		0.33	0.80	0.31
GRU		0.33	0.81	0.31
Bi-GRU		0.38	0.81	0.28
RNN		0.33	0.81	0.31
Bi-RNN		0.32	0.79	0.30
CNN		0.26	0.52	0.21
Transfer Learning				
BERT		0.07	0.43	0.12
XLNet		0.43	0.19	0.26

*__Train Data__: Train-CM-CM-MEC-21 (80%)

*__Test Data__: Test-CM-CM-MEC-21 (20%) + Test-E-CM-MEC-21 (20%) + Test-RU-CM-MEC-21 (20%)

TABLE 7. Results obtained by applying various classical machine learning, deep learning and transfer learning techniques (Exp4).

Features	ML Algorithms	Mi_P	Mi_R	Micro F_1
Classical Machine Learning				
Word 1-gram	**OVR (NB)**	**0.60**	**0.57**	**0.60**
Word 2-grams	OVR (SVC)	0.57	0.52	0.57
Word 3-grams	OVR (RF)	0.53	0.50	0.53
Char 3-grams	OVR (NB)	0.57	0.61	0.57
Char 4-grams	OVR (NB)	0.57	0.59	0.57
Char 5-grams	OVR (NB)	0.57	0.58	0.57
Char 6-grams	OVR (NB)	0.57	0.57	0.57
Char 7-grams	OVR (NB)	0.56	0.55	0.56
Char 8-grams	OVR (NB)	0.55	0.53	0.55
Char 9-grams	OVR (NB)	0.55	0.52	0.55
Char 10-grams	OVR (NB)	0.55	0.52	0.55
Deep Learning				
DL Algorithm		**Mi_P**	**Mi_R**	**Micro Avg. F_1**
LSTM		0.33	0.81	0.31
Bi-LSTM		0.50	0.78	0.29
GRU		0.50	0.78	0.29
Bi-GRU		0.26	0.79	0.27
RNN		0.33	0.81	0.31
Bi-RNN		0.49	0.77	0.28
CNN		0.50	0.59	0.26
Transfer Learning				
BERT		0.09	0.47	0.15
XLNet		0.43	0.26	0.32

*__Train Data__: Train-E-CM-MEC-21 (80%) + Train-RU-CM-MEC-21 (80%) + Train-CM-CM-MEC-21 (80%)

*__Test Data__: Test-E-CM-MEC-21 (20%) + Test-RU-CM-MEC-21 (20%) + Test-CM-CM-MEC-21 (20%)

achieved on a short length of n for both words and character n-grams. A possible reason for lower performance on longer length is that when we generate n-gram features for longer lengths of n, it is likely to capture unwanted, noisy, and irrelevant information to predict multiple emotions. Consequently, the performance of the machine learning algorithms decreases.

Regarding the machine learning algorithms, a combination of OVR multi-label and NB single-label machine learning algorithms performed best (Micro Precision = 0.67, Micro Recall = 0.54, Micro F_1 = 0.67) as compared to other algorithms. One possible reason behind this is that OVR and NB are simple, efficient, and able to handle noisy data. Moreover, OVR and NB do not require a huge dataset to work well.

Regarding the combinations of experiments, in Exp1 and Exp2, there is not much variation in results. A possible reason is that the training model has mixed dataset features, characteristics, and knowledge of all three source languages. Considering Exp3 and Exp4, there is variation in results. This indicates that code-mixed (Train-CM-MEC-21) data is

TABLE 8. Best results obtained on Exp1, Exp2, Exp3, and Exp4.

Experiment No.	Train-Data	Test-Data	Features	ML Algorithms	Micro F_1
Exp1	E-CM-MEC-21 + RM-CM-MEC-21	CM-CM-MEC-21	Word 2-grams	OVR (Bagging)	0.64
Exp2	CM-CM-MEC-21	CM-CM-MEC-21	Char 3-grams	OVR (SVC)	0.66
Exp3	CM-CM-MEC-21	CM-CM-MEC-21 + E-CM-MEC-21 + RU-CM-MEC-21	**Word 1-gram**	**OVR (SVC)**	**0.67**
Exp4	CM-CM-MEC-21 + E-CM-MEC-21 + RU-CM-MEC-21	CM-CM-MEC-21 + E-CM-MEC-21 + RU-CM-MEC-21	Word 1-gram	OVR (NB)	0.60

more suitable in model training compared to a combination of English (monolingual), Roman Urdu (monolingual), and code-mixed (English + Roman Urdu) SMS message on our proposed CM-MEC-21 corpus.

The main findings of these four experiments are: (1) classical machine learning outperforms deep learning and transfer learning methods in all four experiments, (2) overall highest Micro F_1 of 0.67 is obtained (Exp3), indicating that multi-label emotion classification on code-mixed data is a complex task, (3) deep learning and transfer learning methods fail to give promising results when we have small training data, and (4) change in training data also changes the best-performing features in the majority of experiments, i.e., Exp1, Exp2, and Exp3.

To conclude, overall best results (see Table 8) are obtained using state-of-the-art classical machine learning methods with word uni-gram (Micro $F_1 = 0.67$) and OVR multi-label machine learning algorithm when training on code-mixed (Train-CM-CM-MEC-21) and testing by combining code-mixed, Roman Urdu, and English (Test-CM-CM-MEC-21 + Test-E-CM-MEC-21 + Test-RU-CM-MEC-21) SMS messages.

VII. CONCLUSION

Code-mixed (English and Roman Urdu) text is widely used, especially in the South Asian community. However, it is not explored for the multi-label emotion classification problem. As described in this paper, our novel contribution is a newly developed and publicly available benchmark code-mixed and multi-label SMS messages-based corpus for the multi-label emotion classification task. The corpus consists of 11,914 code-mixed multi-label SMS messages manually annotated for the presence/absence of the following 12 emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral (no emotion). In addition to dataset creation, we applied state-of-the-art machine learning (content-based – 3-word n-gram features and eight character n-gram features), deep learning, and transfer learning-based methods for multi-label emotion classification task on our proposed CM-MEC-21 corpus. The best results (see Table 8) obtained using state-of-the-art machine learning methods with word uni-gram (Micro $F_1 = 0.67$) and OVR multi-label machine learning algorithm, when training on code-mixed (Train-CM-CM-MEC-21) and testing by combining code-mixed, Roman Urdu, and English (Test-CM-CM-MEC-21 + Test-E-CM-MEC-21 + Test-RU-CM-MEC-21) multi-label SMS messages.

In the future, we plan to apply other transfer learning-based models such as RoBERTa, DistilBERT, etc., to our proposed corpus. An ensemble of the model would be considered to increase classification performance.

ACKNOWLEDGMENT

The authors utilize the computing resources brought to them by the CONACYT through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico.

REFERENCES

[1] S. Eurobarometer, "Europeans and their languages," European Commission, Brussels, Belgium, Tech. Rep., 2006.
[2] M. Fatima, K. Hasan, S. Anwar, and R. M. A. Nawab, "Multilingual author profiling on Facebook," *Inf. Process. Manage.*, vol. 53, no. 4, pp. 886–904, 2017.
[3] M. A. Ashraf, R. M. A. Nawab, and F. Nie, "Author profiling on bi-lingual tweets," *J. Intell. Fuzzy Syst.*, vol. 39, no. 2, pp. 2379–2389, 2020.
[4] M. H. F. Siddiqui, I. Ameer, A. F. Gelbukh, and G. Sidorov, "Bots and gender profiling on Twitter," in *Proc. Work. Notes Conf. Labs Eval. Forum (CLEF)*, Lugano, Switzerland, Sep. 2019, pp. 1–8.
[5] I. Ameer, M. H. F. Siddiqui, G. Sidorov, and A. Gelbukh, "CIC at SemEval-2019 task 5: Simple yet very efficient approach to hate speech detection, aggressive behavior detection, and target classification in Twitter," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 382–386.
[6] M. Fatima, S. Anwar, A. Naveed, W. Arshad, R. M. A. Nawab, M. Iqbal, and A. Masood, "Multilingual SMS-based author profiling: Data and methods," *Natural Lang. Eng.*, vol. 24, no. 5, pp. 695–724, Sep. 2018.
[7] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 1–17.
[8] A. G. Shahraki and O. R. Zaiane, "Lexical and learning-based emotion mining from text," in *Proc. Int. Conf. Comput. Linguistics Intell. Text Process.*, vol. 9, 2017, pp. 24–55.
[9] C. Huang, A. Trabelsi, X. Qin, N. Farruque, and O. R. Zaïane, "Seq2Emo for multi-label emotion classification based on latent variable chains transformation," 2019, *arXiv:1911.02147*.
[10] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proc. 4th Int. Workshop Semantic Eval. (SemEval)*, 2007, pp. 70–74.
[11] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *Proc. Int. Conf. Text, Speech Dialogue*. Springer, 2007, pp. 196–205.
[12] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," 2017, *arXiv:1710.03957*.
[13] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, "Sentiment, emotion, purpose, and style in electoral tweets," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 480–499, 2015.
[14] S. M. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets," 2017, *arXiv:1708.03696*.
[15] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Detecting emotion stimuli in emotion-bearing sentences," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Springer, 2015, pp. 152–165.
[16] H. Schuff, J. Barnes, J. Mohme, S. Padó, and R. Klinger, "Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus," in *Proc. 8th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2017, pp. 13–23.
[17] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," in *Proc. Hum. Lang. Technol. Conf. Conf. Empirical Methods Natural Lang. Process.*, 2005, pp. 579–586.
[18] S. M. Mohammad, "# Emotional tweets," in *Proc. 1st Joint Conf. Lexical Comput. Semantics, Main Conf. Shared Task, 6th Int. Workshop Semantic Eval.*, vols. 1–2, 2012, pp. 246–255.
[19] C. Liu, M. Osama, and A. de Andrade, "DENS: A dataset for multi-class emotion analysis," 2019, *arXiv:1910.11769*.
[20] C. Baziotis, N. Athanasiou, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos, "NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning," 2018, *arXiv:1804.06658*.

[21] A. Turner. (Jul. 2021). *How Many People Have Smartphones Worldwide*. [Online]. Available: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world
[22] S. L. Videbeck, *Psychiatric Mental Health Nursing*. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2006.
[23] (Mar. 2021). *Anticipation*. [Online]. Available: https://en.wikipedia.org/wiki/Anticipation
[24] C. L. Badour and M. T. Feldner, "The role of disgust in posttraumatic stress: A critical review of the empirical literature," *J. Exp. Psychopathol.*, vol. 9, no. 3, pp. pr–032813, 2018.
[25] (Jul. 2021). *Fear*. [Online]. Available: https://en.wikipedia.org/wiki/Fear
[26] (May 2021). *Joy*. [Online]. Available: https://en.wikipedia.org/wiki/Joy
[27] (Apr. 2021). *Optimism*. [Online]. Available: https://en.wikipedia.org/wiki/Optimism
[28] O. Bennett, *Cultural Pessimism: Narratives of Decline in the Postmodern World*. Edinburgh, U.K.: Edinburgh Univ. Press, 2001.
[29] *Sorrow*. [Online]. Available: https://www.dictionary.com/browse/sorrow
[30] (Mar. 2013). *Surprise*. [Online]. Available: https://simple.wikipedia.org/wiki/Surprise
[31] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, 1995.
[32] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proc. 12th Int. Workshop Semantic Eval.*, New Orleans, LA, USA, 2018, pp. 1–17.
[33] P. S. Bayerl and K. I. Paul, "What determines inter-coder agreement in manual annotations? A meta-analytic investigation," *Comput. Linguistics*, vol. 37, no. 4, pp. 699–725, Dec. 2011.
[34] K. Boland, A. Wira-Alam, and R. Messerschmidt, "Creating an annotated corpus for sentiment analysis of German product reviews," GESIS-Leibniz Inst. Social Sci., Mannheim, Germany, Tech. Rep., 2013.
[35] I. Mozetič, M. Grčar, and J. Smailović, "Multilingual Twitter sentiment classification: The role of human annotators," *PLoS ONE*, vol. 11, no. 5, May 2016, Art. no. e0155036.
[36] E. Öhman, "Emotion annotation: Rethinking emotion categorization," in *Proc. DHN Post*, 2020, pp. 134–144.
[37] B. Thomas, P. Vinod, and K. Dhanya, "Multiclass emotion extraction from sentences," *Int. J. Sci. Eng. Res.*, vol. 5, no. 2, pp. 12–15, 2014.
[38] W. Wang, "Automatic emotion identification from text," Ph.D. dissertation, Wright State Univ., Dayton, OH, USA, 2015.
[39] I. Ameer, N. Ashraf, G. Sidorov, and H. Gómez Adorno, "Multi-label emotion classification using content-based features in Twitter," *Computación y Sistemas*, vol. 24, no. 3, Sep. 2020.
[40] S. Mohammad, X. Zhu, and J. Martin, "Semantic role labeling of emotions in tweets," in *Proc. 5th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2014, pp. 32–41.
[41] I. Ameer, G. Sidorov, and R. M. A. Nawab, "Author profiling for age and gender using combinations of features of various types," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4833–4843, May 2019.
[42] I. Ameer and G. Sidorov, "Author profiling using texts in social networks," in *Handbook of Research on Natural Language Processing and Smart Service Systems*. Hershey, PA, USA: IGI Global, 2021, pp. 245–265.
[43] I. Pervaz, I. Ameer, A. Sittar, and R. M. A. Nawab, "Identification of author personality traits using stylistic features," in *Proc. Eval. Labs Workshop–Work. Notes Papers (CLEF)*, Toulouse, France, Sep. 2015.
[44] I. Pervaz, I. Ameer, A. Sittar, and R. M. A. Nawab, "Identification of author personality traits using stylistic features: Notebook for PAN at CLEF 2015," in *CLEF (Working Notes)*, 2015, pp. 1–7.
[45] H. Yang, J. Han, and K. Min, "A multi-column CNN model for emotion recognition from EEG signals," *Sensors*, vol. 19, no. 21, p. 4736, Oct. 2019.
[46] J. Lee and S. K. Yoo, "Recognition of negative emotion using long short-term memory with bio-signal feature compression," *Sensors*, vol. 20, no. 2, p. 573, Jan. 2020.
[47] R. Rana, "Gated recurrent unit (GRU) for emotion classification from noisy speech," 2016, *arXiv:1612.07778*.
[48] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*.
[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[50] X. Li, X. Fu, G. Xu, Y. Yang, J. Wang, L. Jin, Q. Liu, and T. Xiang, "Enhancing BERT representation with context-aware embedding for aspect-based sentiment analysis," *IEEE Access*, vol. 8, pp. 46868–46876, 2020.
[51] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," 2019, *arXiv:1903.09588*.
[52] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," 2019, *arXiv:1904.02232*.
[53] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of Bert, Roberta, Distilbert, and Xlnet for text-based emotion recognition," in *Proc. 17th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2020, pp. 117–121.

IQRA AMEER received the B.S. degree in computer science from COMSATS University Islamabad, Lahore Campus, Pakistan, the M.S. degree (Hons.) in computer science from the Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico, in 2017, where she is currently pursuing the Ph.D. degree. Her research interests include natural language processing, sentiment analysis, author attribution, and emotion analysis.

GRIGORI SIDOROV received the Ph.D. degree in philology (computational and structural linguistics) from Lomonosov Moscow State University, Russia, in 1996. He is currently a Full Professor with the Natural Language Processing Laboratory, Center for Computing Research, National Polytechnic Institute, Mexico. He is also a National Researcher of Mexico level 3 (highest level, about 2–3% of all researchers in Mexico have this level) and a member of the Mexican Academy of Sciences. His research interests include computational linguistics, natural language processing, and the application of machine learning techniques to NLP tasks. He is the Editor-in-Chief of the research journal *Computación y Sistemas* (WoS, Scopus, and CONACYT), author of more than 270 scientific publications, of which nine books.

HELENA GÓMEZ-ADORNO received the Ph.D. degree in computer science from the Centro de Investigación en Computación, National Polytechnic Institute, Mexico. She is currently an Associate Researcher at the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM. She has worked on question answering systems, semantic similarity, authorship attribution, author profiling, and text classification problems. Her research interest includes natural language processing. She is a member of the National System of Researchers of CONACYT Level 1.

RAO MUHAMMAD ADEEL NAWAB received the Ph.D. degree in computer science from The University of Sheffield, U.K. He is currently working as an Assistant Professor with the Computer Science Department, COMSATS University Islamabad, Lahore Campus, Pakistan. His research interests include natural language processing, text reuse, plagiarism detection, author profiling, word sense disambiguation, text summarization, image captioning, information retrieval, question answering systems, and paraphrase detection.