Findings of WASSA 2023 Shared Task: Multi-Label and Multi-Class Emotion Classification on Code-Mixed Text Messages

Necva Bölücü*[1], Iqra Ameer*[2], Ali Al Bataineh[3], and Hua Xu[4]

[1]Data61, CSIRO, Sydney, Australia
[2]Division of Science and Engineering, Penn State University at Abington, Pennsylvania, USA
[3]Electrical and Computer Engineering Norwich University, USA
[4]Section of Biomedical Informatics and Data Science, Yale School of Medicine, USA

Abstract

We present the results of the WASSA 2023 Shared-Task 2: Emotion Classification on code-mixed text messages (Roman Urdu + English), which included two tracks for emotion classification: multi-label and multi-class. The participants were provided with a dataset of code-mixed SMS messages in English and Roman Urdu labeled with 12 emotions for both tracks. A total of 5 teams (19 team members) participated in the shared task. We summarized the methods, resources, and tools used by the participating teams. We also made the data freely available for further improvements to the task.

1 Introduction

In recent times, the growing number of Internet users and the proliferation of diverse online platforms have led to a significant surge in individuals expressing their opinions and attitudes on government websites, microblogs, and other social media platforms. Consequently, there is growing interest in effectively extracting people's sentiments and emotions towards events from such subjective information. To address this, *Natural Language Processing* (NLP) employs emotion analysis called *Emotion Classification*. Emotion Classification is one of the most challenging NLP tasks, in which a given text is assigned to the most appropriate emotion(s) that best reflect the author's mental state of mind (Tao and Fang, 2020), where emotions can be anger, joy, sadness, surprise, etc. People freely express their feelings, arguments, opinions, and thoughts on social media. Therefore, this task plays a pivotal role in uncovering valuable insights from user-generated content, and more and more attention is being paid to automatic tools for classifying users' emotion(s) from written text. Emotion classification has applications in several domains, including financial marketing (Zhang et al., 2016; Yang et al., 2020; Lysova and Rasskazova, 2019), medicine (Lin et al., 2016; Saffar et al., 2022; Huang et al., 2023), education (Huang and Zhang, 2019; Zhang et al., 2020b; Carstens et al., 2019), etc.

There are two different views on the classification of emotions. Ameer et al. (2020) stated that emotions are dependent; one emotional expression can be linked to multiple emotions (Deng and Ren, 2020). Therefore, the emotion classification problem should be defined as *Multi-Label Emotion Classification* (MLEC). MLEC is the task of assigning all possible emotions for a written text that best presents the author's mental state. The other view is that written data is associated with only one emotion (Ameer et al., 2022), which defines the problem as a *Multi-class Emotion Classification* (MCEC) problem. MCEC is the task of assigning one most dominating emotion to the given piece of text that best represents the mental state of an author.

In this paper, we present the WASSA 2023 Shared Task: Multi-Label and Multi-Class Emotion Classification on Code-Mixed Text Messages. We used the same dataset provided by (Ameer et al., 2022) composed of code-mixed (English + Roman Urdu) SMS messages originally collected for MLEC. Each SMS message is annotated for the absence/presence of 12 multiple emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral (no emotion)) provided by SemEval-2018 Task 1: Affect in Tweets (Mohammad et al., 2018) (see Section 3 for more details). The shared task consists of two tracks:

- Track 1 - MLEC: The formulation of this track is to predict all possible emotion labels from code-mixed SMS messages.
- Track 2 - MCEC: The formulation of this track is to predict a single most dominating emotion

[1]*Necva Bölücü and Iqra Ameer contributed equally to this work.

from code-mixed SMS messages.

7 teams participated in this shared task: 3 teams submitted results to *MLEC* and 7 teams submitted results to *MCEC* tracks[1]. The tracks were designed using CodaLab[2], allowing teams to submit one official result during the evaluation phase and multiple results during the training phase. During the evaluation phase, each team was allowed to submit their results by a certain deadline, after which the final submission was considered for ranking. The best result for *Track 1 - MLEC* was Multi-Label Accuracy = 0.9782, and the best result for *Track 2 - MCEC* was Macro F_1 = 0.9329.

The rest of the paper is structured as follows: Section 2 provides an overview of related work. Section 3 presents the details of the datasets for both tracks. The task description is outlined in Section 4, while the official results are presented in Section 5. Section 6 provides a discussion of the various systems that participated in both tracks. Finally, our work is concluded in Section 7.

2 Related Work

In recent years, extensive research has been conducted on emotion classification (Ren et al., 2017; Tang et al., 2019; Zhang et al., 2020a). Among supervised machine learning techniques, Random Forest, Logistic Regression, Naïve Bayes, Support Vector Machine, Bagging, AdaBoost, and Decision Tree are widely used for emotion classification problems (Ameer et al., 2020, 2022; Hadwan et al., 2022; Edalati et al., 2022).

The success of deep learning models in various NLP tasks, including Neural Machine Translation (NMT) (Wang et al., 2017; Song et al., 2019) and Semantic Textual Similarity (STS) (Wu et al., 2021; Zhang and Lan, 2021), has led them to be applied to the emotion classification problem as well. Notably, deep learning models, LSTM (Baziotis et al., 2018; Gee and Wang, 2018), CNN (Kim et al., 2018), GRU (Eisner et al., 2016; Alswaidan and Menai, 2020), GNN (Ameer et al., 2023b) and Transformers (e.g., BERT, XLNet, DistilBERT, and RoBERTa) (Ameer et al., 2020; Ding et al., 2020; Ameer et al., 2022, 2023a) have been utilized in this context.

There have been several efforts in the literature to construct benchmark corpora for emotion classification tasks (Illendula and Sheth, 2019; Demszky et al., 2020; Xu et al., 2015; Saputra et al., 2022; Ashraf et al., 2022; Ilyas et al., 2023). However, the existing efforts have primarily focused on monolingual datasets. In particular, SemEval has organized a number of international competitions (Mohammad et al., 2018; Strapparava and Mihalcea, 2007) that have published monolingual benchmark corpora for MLEC, which serve as valuable resources for developing, comparing, and evaluating approaches. Regarding the code-mixed task, a few benchmark corpora have been developed for MLEC (Vijay et al., 2018; Sinha et al., 2021; Sasidhar et al., 2020; Lee and Wang, 2015; Tan et al., 2020; Plaza-del Arco et al., 2020).

Vijay et al. (2018) developed a Hindi-English code-mixed corpus by collecting 2,866 tweets from the past eight years. The corpus was annotated with Ekman's six emotion labels, including anger, disgust, fear, happiness, sadness, and surprise. Each tweet in the corpus was labeled with its source language and the causal language of the expressed emotion. Another effort by Sinha et al. (2021) involved the development of a Hindi-English code-mixed corpus of 15,997 Facebook status updates. These updates were annotated with emotions such as joy, sadness, anger, fear, trust, disgust, surprise, anticipation, and love. Similarly, Sasidhar et al. (2020) created a Hindi-English code-mixed corpus for single-label emotion classification. This corpus consisted of 12,000 texts gathered from Twitter, Instagram, and Facebook posts. It was manually annotated with three basic emotion labels: happy, sad, and anger.

For Chinese-English code-mixed corpora, Lee and Wang (2015) compiled a multilingual corpus by collecting code-switching data from `Weibo.com`, a popular Chinese social networking website. The corpus contained 2,313 posts annotated with five basic emotions: anger, fear, happiness, sadness, and surprise. The posts covered various domains such as life, finance, service, celebrities, products, and politics, with happiness being the most dominant emotion.

In the context of Malaysian code-mixed corpora, Tan et al. (2020) developed a large Twitter corpus consisting of 295,817 Tweets in the Malaysian language (Malay, Malaysian slang, and English). The corpus was annotated with six basic emotion

[1]Only 5 of the teams submitted system description papers.

[2]Details of task descriptions, datasets, and results are in CodaLab https://codalab.lisn.upsaclay.fr/competitions/10864

classes: anger, fear, happiness, love, sadness, and surprise. Additionally, Plaza-del Arco et al. (2020) compiled a multi-label and code-mixed emotion corpus based on events in April 2019. The corpus included 7,303 English tweets and 8,409 Spanish tweets. Each tweet was assigned one of Ekman's fundamental emotions, such as anger, surprise, disgust, enjoyment, fear, and sadness, or labeled as neutral or other emotions.

While existing code-mixed corpora mainly focused on English combined with Spanish, Malaysian, Hindi, and other languages for tweets, a benchmark code-mixed (English + Roman Urdu) dataset with proposed models to solve the problem for the MLEC task was lacking. To address this gap, the code-mixed dataset developed by Ameer et al. (2022) for MLEC was used for the shared task by extending the problem for MLEC and MCEC problems.

3 Dataset Compilation Process

The dataset–CM-MEC-21 corpus–utilized for the shared task is developed for the MLEC task and consists of code-mixed (English + Roman Urdu) SMS messages (Ameer et al., 2022). In this section, we first provide the details of the original dataset and then describe the dataset preparation process for the *MCEC* track of the shared task.

The dataset contains code-mixed (English + Roman Urdu) SMS messages which are manually selected from SMS-AP-18 corpus (Fatima et al., 2018) and annotated by three annotators for the presence/absence of 12 emotions as in SemEval-2018 (Mohammad et al., 2018) for the MLEC task. Therefore, we used the dataset for the MLEC track of the shared task since it is already annotated for the MLEC using a set of 12 emotions: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral (no emotion).

For the MCEC track, the annotators annotated each code-mixed (English + Roman Urdu) SMS message with the most dominating emotion among all the labels assigned for MLEC. In cases where a code-mixed SMS message did not convey any particular emotion, only the "neutral" label was assigned.

We randomly split the MLEC and MCEC track datasets into train (80%), development (10%), and test (10%) sets. Table 1 represents the train, development, and test splits. The distributions of emotions for MLEC and MCEC tracks for each set are presented in Tables 2 and 3, respectively. The dataset used in the shared task is publicly available[3].

Track	Train	Dev	Test	Total
MLEC	9530	1191	1191	11912
MCEC	9530	1191	1191	11912

Table 1: Statistical details of train, development, and test set for MLEC and MCEC tracks.

Emotion	Train	Dev	Test
Anger	271	41	35
Anticipation	1046	135	134
Disgust	955	134	124
Fear	522	58	51
Joy	1213	144	142
Love	265	34	34
Neutral	3247	404	394
Optimism	1065	133	121
Pessimism	219	26	29
Sadness	638	65	85
Surprise	281	27	34
Trust	1185	145	160

Table 2: Distribution of emotion labels in the MLEC track.

Emotion	Train	Dev	Test
Anger	226	35	26
Anticipation	832	94	97
Disgust	687	113	98
Fear	453	52	55
Joy	1022	131	123
Love	187	17	24
Neutral	3262	388	399
Optimism	880	110	103
Pessimism	178	29	35
Sadness	486	62	69
Surprise	199	35	28
Trust	1118	125	134

Table 3: Distribution of emotion labels in the MCEC track.

[3] https://github.com/wassa23codemixed/codemixed

4 Task Description

We set up the tracks in CodaLab[4]. Section 4.1 describe the tracks of the shared task and dataset, resources, and evaluation metrics are explained in Section 4.2.

4.1 Tracks

Track 1 - Multi-Label Emotion Classification (MLEC): The problem of this task is to classify each code-mixed SMS message as "neutral or no emotion" or as one or more of eleven given emotions (anger, anticipation, disgust, fear, joy love, optimism, pessimism, sadness, surprise, trust) that best represent the mental state of the author.

Track 2 - Multi-Class Emotion Classification (MCEC): The problem of this task is to predict an emotion label from the emotion set, as well as *no emotion tag (neutral)* for each code-mixed SMS message.

4.2 Setup

Dataset: Participants are provided with the dataset described in Section 3. Participants are allowed to use external datasets in the training phase or use data augmentation techniques to improve their systems.

Team	Accuracy	Micro F_1	Macro F_1
YNU-HPCC	**0.9782**	**0.9854**	**0.9869**
CTcloud	0.9723	0.9815	0.9833
wsl&zt	0.9110	0.9407	0.9464
baseline	0.7321	0.8514	0.8347

Table 4: Results of the teams participating in the MLEC track.

Emotion	YNU-HPCC	CTcloud	wsl&zt
Anger	86.67	97.80	80.00
Anticipation	90.49	88.81	81.69
Disgust	95.00	95.54	93.12
Fear	97.67	96.00	94.36
Joy	92.13	98.97	86.83
Love	91.97	90.70	90.00
Optimism	96.46	88.95	82.44
Pessimism	89.25	80.00	84.55
Sadness	95.17	98.91	95.75
Surprise	93.33	97.19	97.40
Trust	85.90	85.43	85.44

Table 5: Class-wise MLEC results (*100) of the teams participating in the MLEC track.

[4] https://codalab.lisn.upsaclay.fr/competitions/10864

Resources and Systems Restrictions: The organizers allowed participants to use any third-party tools, lexical resources, additional train data, or synthetic datasets generated by AI models for the tasks, nor did they apply any restrictions on the participants.

System Evaluation: The official competition evaluation script for MLEC was multi-label accuracy (or Jaccard index), and Macro F_1 was used for MCEC. In addition to the official evaluation metrics, Micro and Macro F_1 scores for MLEC and Accuracy, Macro Precision, and Macro Recall for MCEC were also used as secondary evaluation metrics to provide a different perspective on the results.

5 Results and Discussion

5.1 Multi-Label Emotion Classification

Table 4 presents the main results for the MLEC track. 3 teams submitted their results (2 of them submitted their papers). *YNU-HPCC* ranked first in MLEC track (Multi-label Accuracy = 0.9782), which is very close to team *CTcloud* (Multi-label Accuracy = 0.9723), which ranked second. Table 5 provides the class-wise Macro F_1 results for the teams participating in the MLEC track.

5.2 Multi-Class Emotion Classification

Table 6 presents the main results for the MCEC track. 7 teams submitted their results (5 of them submitted their system description papers), and the best-performing team was *YNU-HPCC* (Macro F_1 = 0.9329).

We also provided class-wise Macro F_1 results of the teams participating in the MCEC track in Table 7 to get more insights. Due to the high frequency in the training set of the dataset, the submitted systems achieved higher Macro F_1 scores for Neutral, Trust, Joy, and Optimism labels compared to other emotion labels.

6 Summary of Participating Systems

WASSA 2023 Shared Task on Multi-Label and Multi-Class Emotion Classification on Code-Mixed Text Messages received 5 system description papers. The results of the systems are represented in Tables 4 and 6 for MLEC and MCEC tracks, respectively. Only two five systems attempted the MLEC and MCEC tasks, while the others did not submit results for the MLEC task.

Team	Macro F_1	Accuracy	Macro Precision	Macro Recall
YNU-HPCC	**0.9329**	**0.9488**	**0.9488**	**0.9488**
CTcloud	0.8917	0.9219	0.9219	0.9219
wsl&zt	0.7359	0.7699	0.7699	0.7699
anedilko	0.7038	0.7313	0.7313	0.7313
baseline	0.7014	0.7298	0.7298	0.7298
PrecogIIITh	0.6061	0.6734	0.6734	0.6734
BpHigh	0.3764	0.5642	0.5642	0.5642

Table 6: Results of the teams participating in the MCEC track.

Emotion	YNU-HPCC	CTcloud	wsl&zt	anedilko	PrecogIIITh	BpHigh
Anger	90.20	80.85	66.67	65.45	53.06	0.00
Anticipation	92.55	87.70	68.11	56.11	58.88	35.64
Disgust	91.63	89.20	67.80	69.32	57.00	37.17
Fear	96.49	91.07	75.73	75.25	60.00	26.53
Joy	94.17	93.23	88.16	80.00	82.20	82.03
Love	91.30	77.27	75.00	72.34	57.89	45.71
Neutral	97.48	95.31	80.09	79.14	73.15	71.93
Optimism	94.34	93.72	74.37	70.94	67.94	58.45
Pessimism	94.29	93.94	67.80	67.69	55.56	0.00
Sadness	94.96	91.04	75.71	77.61	67.16	42.67
Surprise	97.27	84.00	65.22	60.87	28.57	0.00
Trust	94.81	92.72	78.46	69.80	65.93	51.56

Table 7: Class-wise MCEC results (*100) of the teams participating in the MCEC track.

Technique / Model	Submission Count
BERT	1
MBERT	1
RoBERTa	1
XLM-RoBERTa	3
IndicBERT	1
MuRIL	1
XGBClassifier	1
Prompt Tuning	1
Prompt Engineering	1

Table 8: Summary of techniques and architectures used in submissions.

6.1 Machine Learning Architectures

All systems submitted results to the shared task applied deep learning models for MLEC and MCEC tracks. Table 8 provides a high-level summary of the frequency of architectures and techniques used by multiple systems. There are similarities between the four systems based on transformer-based language models. One system deviated from the others using ChatGPT with prompt tuning for the shard task tracks. Three of the systems applied pre-processing (using an emoticon dictionary (*CTcloud*), English translation of code-mixed sentences using ChatGPT (*PrecogIIITh*), converting multi-class labels to multi-label labels with one hot encoding (*YNU-HPCC*)). Only one of the systems used data augmentation in the training phase (*BpHigh*).

With increasing attention to prompt tuning and prompt engineering for extracting knowledge from language models, two of the five systems attempted prompt tuning and engineering for the tasks.

6.2 Features and Resources

For a given code-mixed text, emotion(s) classification is a challenging task in the NLP domain. Teams were allowed to use external resources, which can be data, a lexicon, or contextual embeddings that can improve the performance of systems. Table 9 provides the details of features and resources used in the submitted system description papers.

The emotion lexicon is created by gathering the icons in the training set and collecting more

Features	# of team	MLEC	MCEC
Emotion lexicon	1	✓	✓
ChatGPT	2		✓
External dataset	1		✓
Framework	2	✓	✓

Table 9: Features and resources used in the submitted system description papers.

icons from the Internet [5] (*CTcloud*). ChatGPT is used in the submitted system description papers for translation (*PrecogIIITH*) and prompt engineering (*anedilko*).

Moreover, participants used external datasets in the shared task, such as HS-RU-20 (Khan et al., 2021), Roman Urdu Hate Speech (Rizwan et al., 2020), and Hing-Corpus (Nayak and Joshi, 2022). These datasets are used to train the transformer model with contrastive learning (*BpHigh*).

SetFit[6] (Tunstall et al., 2022) (*BpHigh*) and OpenPrompt[7] (Ding et al., 2021) (*CTcloud*) are used as frameworks in the systems. While SetFit is a framework to build a robust sentence classifier for small datasets that helps finetune sentence transformers on the dataset with contrastive learning, Openprompt is a framework to adapt pre-trained language models (PLMs) to downstream NLP tasks.

6.3 System Specifies

YNU-HPCC, the team ranked first, developed a model using a hybrid dataset approach–combined MLEC and MCEC datasets with a unified multilingual pre-trained model. They applied preprocessing step in the training phase to convert multi-class labels to multi-label labels with one hot encoding. They applied Kullback-Leibler (KL) (Eguchi and Copas, 2006) to obtain mixed annotation labels, combining two tracks and fine-tuning XLM-RoBERTA (Conneau et al., 2019). In inference, they separately obtained the results for two tracks with fine-tuned XLM-RoBERTa.

CTcloud, the team ranked second, applied preprocessing before the training phase, mapping emoticons to textual form using icon-emotion and Unicode-short name mapping to leverage their rich emotional information for the problem. They applied prompt tuning with zero-shot and few-shot approaches for GPT-3. They also applied soft-prompt following Zhu et al. (2022) with manual and soft verbalizer using XLM-RoBERTa (Conneau et al., 2019). The best results are obtained with soft prompts and soft verbalizers. They built their system using OpenPrompt (Zhu et al., 2022). In the experiments, they test base and large versions of XLM-RoBERTa as well as the fine-tuned XLM-ROBERTa for the problem. It is found that when the fine-tuned model is used, only a small amount of prompt tuning is required to obtain satisfactory results. On the other hand, XLM-RoBERTa requires more prompt tuning.

anedilko developed a system for MCEC track with prompt engineering on Chat-GPT API. For the prompts, they chose 100 samples from the training set in terms of the cosine similarity of the samples in the training and development sets using embedding API[8]. They also apply XGB Classifier (Chen and Guestrin, 2016), which used character n-grams as features as the baseline model.

PrecogIIITh fine-tuned multi-lingual transformer-based models, XLM-RoBERTa (Conneau et al., 2019) and IndicBERT (Doddapaneni et al., 2022) for MCEC track. As a third experiment, they used ChatGPT interface[9] to translate code-mixed sentences into English and fine-tuned XLM-RoBERTa with the translated sentences.

BpHigh applied SimCSE (Gao et al., 2021), which uses contrastive learning to obtain sentence embeddings using MuRIL–a transformer-based BERT architecture that supports 17 Indic languages, including English. To train SimCSE, they combined 3 datasets, such as HS-RU-20 (Khan et al., 2021), Roman Urdu Hate Speech (Rizwan et al., 2020), and Hing-Corpus dataset (Nayak and Joshi, 2022).

Table 10 presents the details of the submitted systems to the shared task.

7 Conclusions

This paper presents a shared task on multi-label and multi-class emotion classification for code-mixed (English and Roman Urdu) SMS messages. We provide a comprehensive overview of the task, including its design, data, evaluation process, results, and participating systems. Through the analysis of the systems, we find that most of them employ fine-tuned pre-trained language models for the task

[5] https://en.wikipedia.org/wiki/List$_of_emoticons$ Lastvisited : 06 − 08 − 2023.

[6] https://github.com/huggingface/setfit Last visited: 06-08-2023.

[7] https://github.com/thunlp/OpenPrompt Last visited: 06-08-2023.

[8] https://platform.openai.com/docs/guides/embeddings

[9] https://openai.com/blog/chatgpt

Team Name	# of Authors	MCEC	MLEC	Algorithm
YNU-HPCC	5	✓	✓	Finetune PLM
CTcloud	5	✓	✓	Prompt Tuning
wsl&zt	-	✓	✓	
anedilko	1	✓		Prompt Engineering
Arenborg	-	✓		Finetune PLM
PrecogIIITh	4	✓		Finetune PLM
BpHigh	1	✓		Finetune PLM & Contrastive Learning

Table 10: Summary of all the teams that reported their results

of multi-class emotion classification. While these models have shown success in this domain, our observations indicate the need for additional information to fully leverage their potential. Furthermore, prompt tuning emerges as a prominent area of research, holding great promise for multi-label and multi-class classification tasks, particularly in the context of code-mixed datasets and challenging domains like emotion classification. Finally, prompt engineering emerges as an area that demands further investigation to effectively address the challenges posed by these problems.

References

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, pages 1–51.

Iqra Ameer, Noman Ashraf, Grigori Sidorov, and Helena Gómez Adorno. 2020. Multi-label emotion classification using content-based features in twitter. *Computación y Sistemas*, 24(3):1159–1164.

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023a. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Iqra Ameer, Necva Bölücü, Grigori Sidorov, and Burcu Can. 2023b. Emotion classification in texts over graph neural networks: Semantic representation is better than syntactic. *IEEE Access*.

Iqra Ameer, Grigori Sidorov, Helena Gomez-Adorno, and Rao Muhammad Adeel Nawab. 2022. Multi-label emotion classification on code-mixed text: Data and methods. *IEEE Access*, 10:8779–8789.

Noman Ashraf, Lal Khan, Sabur Butt, Hsien-Tsung Chang, Grigori Sidorov, and Alexander Gelbukh. 2022. Multi-label emotion classification of Urdu tweets. *PeerJ Computer Science*, 8:e896.

Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning.

Alta Carstens, Vanessa Hockly, Maria Petronella Koen, and Elizabeth Johanna Pretorius. 2019. An investigation into the use of emotional intelligence for learning analytics. *International Journal of Educational Technology in Higher Education*, 16(1):1–15.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jie Deng and Feng Ren. 2020. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. *IEEE Transactions on Affective Computing*, 11(3):360–373.

Fei Ding, Xin Kang, Shun Nishide, Zhijin Guan, and Fuji Ren. 2020. A fusion model for multi-label emotion classification based on BERT and topic clustering. In *International Symposium on Artificial Intelligence and Robotics 2020*, volume 11574, pages 98–111. SPIE.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. IndicXTREME: A Multi-Task Benchmark For Evaluating Indic Languages. *arXiv preprint arXiv:2212.05409*.

Maryam Edalati, Ali Shariq Imran, Zenun Kastrati, and Sher Muhammad Daudpota. 2022. The potential of machine learning algorithms for sentiment classification of students' feedback on MOOC. In *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3*, pages 11–22. Springer.

Shinto Eguchi and John Copas. 2006. Interpreting kullback–leibler divergence with the neyman–pearson lemma. *Journal of Multivariate Analysis*, 97(9):2034–2040.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning Emoji Representations from their Description.

Mehwish Fatima, Saba Anwar, Amna Naveed, Waqas Arshad, Rao Muhammad Adeel Nawab, Muntaha Iqbal, and Alia Masood. 2018. Multilingual SMS-based author profiling: Data and methods. *Natural Language Engineering*, 24(5):695–724.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Grace Gee and Eugene Wang. 2018. psyML at SemEval-2018 Task 1: Transfer learning for sentiment and emotion analysis. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 369–376.

Mohammed Hadwan, Mohammed Al-Sarem, Faisal Saeed, and Mohammed A Al-Hagery. 2022. An improved sentiment classification approach for measuring user satisfaction toward governmental services' mobile apps using machine learning methods with feature engineering and smote technique. *Applied Sciences*, 12(11):5547.

Chih-Wei Huang, Bethany CY Wu, Phung Anh Nguyen, Hsiao-Han Wang, Chih-Chung Kao, Pei-Chen Lee, Annisa Ristya Rahmanti, Jason C Hsu, Hsuan-Chia Yang, and Yu-Chuan Jack Li. 2023. Emotion recognition in doctor-patient interactions from real-world clinical video database: Initial development of artificial empathy. *Computer Methods and Programs in Biomedicine*, 233:107480.

Yali Huang and Jinhua Zhang. 2019. A study of the effectiveness of emotion recognition for student learning outcomes in e-learning environments. *Interactive Learning Environments*, 27(7):1019–1032.

Anurag Illendula and Amit Sheth. 2019. Multimodal emotion classification. In *companion proceedings of the 2019 world wide web conference*, pages 439–449.

Abdullah Ilyas, Khurram Shahzad, and Muhammad Kamran Malik. 2023. Emotion Detection in Code-Mixed Roman Urdu-English Text. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–28.

Muhammad Moin Khan, Khurram Shahzad, and Muhammad Kamran Malik. 2021. Hate speech detection in roman urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–19.

Yanghoon Kim, Hwanhee Lee, and Kyomin Jung. 2018. AttnConvnet at SemEval-2018 Task 1: Attention-based Convolutional Neural Networks for Multi-label Emotion Classification.

Sophia Lee and Zhongqing Wang. 2015. Emotion in code-switching texts: Corpus construction and analysis. In *Proceedings of the Eighth SIGHAN workshop on chinese language processing*, pages 91–99.

Kai Lin, Fuzhen Xia, Wenjian Wang, Daxin Tian, and Jeungeun Song. 2016. System Design for Big Data Application in Emotion-Aware Healthcare. *IEEE Access*, 4:6901–6909.

Ekaterina I Lysova and Elena J Rasskazova. 2019. Emotions in financial decision-making: A systematic review. *Journal of Behavioral and Experimental Finance*, 24:100–113.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A Code Mixed Hindi-English Dataset and BERT Language Models. *arXiv preprint arXiv:2204.08398*.

Flor Miriam Plaza-del Arco, Carlo Strapparava, L Alfonso Urena Lopez, and M Teresa Martín-Valdivia. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498.

Han Ren, Yafeng Ren, Xia Li, Wenhe Feng, and Maofu Liu. 2017. Natural logic inference for emotion detection. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 424–436. Springer.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in roman Urdu. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2512–2522.

Alieh Hajizadeh Saffar, Tiffany Katharine Mann, and Bahadorreza Ofoghi. 2022. Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, page 104258.

Karen Etania Saputra, Galih Dea Pratama, Andry Chowanda, et al. 2022. Emotion dataset from Indonesian public opinion. *Data in Brief*, 43:108465.

T Tulasi Sasidhar, B Premjith, and KP Soman. 2020. Emotion detection in hinglish (hindi+ english) code-mixed social media text. *Procedia Computer Science*, 171:1346–1352.

S Sinha, K Saxena, and N Joshi. 2021. Detecting Multi-label emotions from code-mixed Facebook Status Updates. *Indian Journal of Science and Technology*, 14(31):2542–2549.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 70–74.

Kathleen Swee Neo Tan, Tong Ming Lim, and Yee Mei Lim. 2020. Emotion analysis using self-training on Malaysian code-mixed Twitter data. In *International Conferences ICT, Society, and Human Beings*, pages 181–188.

Donglei Tang, Zhikai Zhang, Yulan He, Chao Lin, and Deyu Zhou. 2019. Hidden topic–emotion transition model for multi-level social emotion detection. *Knowledge-Based Systems*, 164:426–435.

Jun Tao and Xiaohui Fang. 2020. Toward multi-label sentiment analysis: a transfer learning based approach. *J. Big Data*, 7(1):10.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient Few-Shot Learning Without Prompts. *arXiv preprint arXiv:2209.11055*.

Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Corpus creation and emotion prediction for Hindi-English code-mixed social media text. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: student research workshop*, pages 128–135.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. ESimCSE: Enhanced Sample Building Method for Contrastive Learning of Unsupervised Sentence Embedding. *CoRR*, abs/2109.04380.

Hua Xu, Weiwei Yang, and Jiushuo Wang. 2015. Hierarchical emotion classification and emotion component analysis on Chinese micro-blog posts. *Expert systems with applications*, 42(22):8745–8752.

Yichi Yang, Yang Liu, and Jian Mao. 2020. Sentiment analysis of financial news and its impact on stock price movements. *Journal of Finance and Data Science*, 6(3):310–320.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020a. Multi-modal Multi-label Emotion Detection with Modality and Label Dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3584–3593.

Junlei Zhang and Zhenzhong Lan. 2021. S-SimCSE: Sampled Sub-networks for Contrastive Learning of Sentence Embedding. *CoRR*, abs/2111.11750.

Lei Zhang, Jian Yang, and Zhiming Tang. 2020b. Designing emotional intelligent tutors: A systematic review of affective computing in education. *Journal of Educational Computing Research*, 57(8):2133–2166.

Shuo Zhang, Yuhong Li, Yan Liu, and Hsinchun Chen. 2016. Emotional advertising: A study of the emotional impact of advertising on consumer behavior. *Journal of Financial Services Marketing*, 21(4):288–299.

Yi Zhu, Xinke Zhou, Jipeng Qiang, Yun Li, Yunhao Yuan, and Xindong Wu. 2022. Prompt-learning for short text classification. *arXiv preprint arXiv:2202.11345*.