

Cost-sensitive multi-label classification in BERT-based models: An inverse-frequency re-weighting approach to emoji pragmatics

Fahad Al Hussen

King Saud University

Mohammed Shormani

Shormani@ibbuniv.edu.ye

Ibb University

Research Article

Keywords: Cost-sensitivity, multi-label classification, inverse-frequency re-weighting, AraBERT, MARBERT, self-attention, multi-head attention, macro/micro functions, digital pragmatics, Facebook

Posted Date: September 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10742668/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

This article studies cost-sensitive multi-label classification (CSMLC) of emoji pragmatic functions in Arabic digital discourse (ADD). It employs 2 Arabic BERT-based models, AraBERT and MARBERT. Beyond sentiment and emotion analysis, the study examines 2 macro pragmatic functions, Affective and Interactional, and 10 micro functions in 8390 unique diglossic Facebook posts containing emojis. To address severe label imbalance, an inverse-frequency cost-sensitive reweighting (IFRW) approach was employed to both models and compared with baseline architectures. We introduced 2 algorithms for finetuning and optimizing both models. The models were evaluated using macro- and micro-averaged precision, recall, and F1, together with Jaccard Accuracy/Score and Hamming Loss. Findings unveil that combining BERT Transformer representations with IFRW improves balanced recognition of underrepresented pragmatic functions. MARBERT consistently outperformed AraBERT with Macro-F1 increasing from 0.53 to 0.75 under reweighting, compared to an increase from 0.50 to 0.54 for AraBERT. MARBERT-ReW also substantially improved recognition of rare functions including *surprise* F1 = 0.73, *agreement* 0.56, *anger* 0.37, and *mitigation* 0.19, all of which received F1 = 0.00 in the baseline models. The study contributes to Arabic NLP by extending emoji analysis beyond sentiment and emotion towards computational modeling of pragmatic functions in ADD.

1. Introduction

Deep learning (DL) has transformed NLP by enabling computational models to learn complex representations from data rather than relying primarily on manually engineered linguistic features (see e.g., Ren et al., 2018; Goodfellow et al., 2016). This development has been particularly important for text classification, where (hierarchical/linear) contextual representations can support the prediction of increasingly fine-grained linguistic and discourse phenomena. In Arabic NLP, however, much of the existing classification research has revolved around sentiment, emotion, and sarcasm, often framing the task in terms of single-label categories or positive–negative polarity. Such formulations do not fully capture the communicative complexity of Arabic digital discourse (ADD), where a single post may simultaneously realize several pragmatic functions through the interaction of linguistic context and emoji use (Yus, 2025; Shormani, 2026). This limitation is particularly relevant to digital pragmatics, where emojis can embody macro pragmatic functions, namely Affective and Interactional, and micro functions, which are both not reducible to sentiment polarity or discrete emotion categories. A single post may simultaneously express Affective function, which in turn may include *sadness*, *love*, and *prayer*, for instance. Another post may contain Interactional function which comprises *humor*, *emphasis*, and *sarcasm*. Such emoji pragmatic uses make cost-sensitive multi-label classification (CSMLC) more appropriate than conventional SLC because multiple non-mutually-exclusive functions can be assigned to the same instance. However, applying CSMLC to Arabic digital pragmatic phenomena constitutes a substantial gap.

More importantly, modeling CSMLC of emoji pragmatics is perhaps untouched, specifically within inverse-frequency re-weighting (IFRW) approach, to the best of our knowledge. Thus, there is a critical gap, which needs to be bridged. Specifically, this study addresses this research problem by examining whether Transformer-based models (TBMs) can learn CSMLC within IFRW, modeling multiple co-occurring

pragmatic functions in ADD and improving their sensitivity to underrepresented functions. We compare 2 Arabic-based TBMs, AraBERT and MARBERT, whose different pretraining profiles provide a particularly relevant basis for comparison in ADD (Antoun et al., 2020; Abdul-Mageed et al., 2022). AraBERT was developed as (Modern) Standard Arabic (SA)-oriented pretrained on formal corpora, but MARBERT was specifically pretrained on dialectal Arabic (DA) large-scale social-media data and is designed to accommodate dialectal and noisy user-generated language (Antoun et al., 2020; Abdul-Mageed et al., 2022). This distinction is central to ADD, or, otherwise, online communication on social media platforms, which is characterized by informal language, dialect mixing, non-standard orthography, spelling variation, and extensive emoji use (Habash, 2010). We introduce 2 DL algorithms to capture MLC emoji pragmatics, and develop IFRW, thus developing a new DL approach to improving digital pragmatics modeling, beyond mere sentiment and emotion single-label classification (SLC). 8390 unique Facebook posts collected via Python are split into 3 mutually exclusive imbalanced datasets, training, validation, and evaluation. The corpus is hierarchically annotated into 2 macro-pragmatic domains, Affective and Interactional, with 10 fine-grained micro-functions constituting the actual prediction targets.

Thus, this article is structured as follows. Section 2 outlines the theoretical foundations and reviews the relevant literature on CSMLC, IFRW, BERT-based models, and Arabic diglossia. Section 3 spells out the study methodology, corpus, annotation scheme, models, re-weighting, optimization, and finetuning Algorithms. Section 4 presents the experimental results and discusses the findings. Section 6 concludes the article and outlines its implications, limitations, and directions for future research.

2. Theoretical foundations and literature review

2.1. Cost-sensitive multi-label classification

TBMs have substantially advanced text classification by replacing extensive manual feature engineering with contextual hierarchical representation learning (Vaswani et al., 2017; Rahali & Akhloufi, 2023; Shormani, 2025), specifically with the advent Artificial Neural Networks (ANNs), and Neural Networking Algorithms (NNAs), connecting this to “how learning happens or could happen in the brain” (Goodfellow et al., 2016, p. 14). In MLC, these contextual representations are mapped through a task-specific classification layer to multiple output labels, allowing several non-mutually-exclusive labels to be assigned to the same input (Tsoumakas & Katakis, 2007; Zhang & Zhou, 2014). This formulation is particularly appropriate for discourse phenomena in which multiple functions may co-occur within the same context (see e.g., Bian, 2023; Huang et al., 2021; Xu et al., 2025; Shormani et al., 2026). In a Transformer-based MLC architecture, the pretrained encoder generates contextualized representations through self-attention, which are subsequently passed to a classification layer producing one independent logit for each target label. The contextual representation generated for the entire input sequence is then passed to a task-specific classification head. The linear layer in multi-label learning generates one independent output score for each label, which distinguishes it from SLC, where the model predicts a single class from mutually exclusive categories (Morozov et al., 2024). To transform these output scores into probabilities, a sigmoid activation function is applied independently to every output neuron (see e.g., Narayan, 1997; Goodfellow et al., 2016; Morozov et al., 2024).

In this study, we conceptualize BERT-based models as deep neural architectures that learn contextual hierarchical representations from data through self-attention and multi-head attention, which are subsequently adapted through task-specific finetuning for CSMLC (cf. Hou et al., 2025; Wang et al., 2026). We further conceptualize CSMLC as a supervised learning framework in which a single input instance can be assigned multiple non-mutually-exclusive labels, while the learning objective assigns greater importance to errors involving underrepresented labels (cf. Rian, 2023; Conneau et al., 2020). We introduce CSMLC algorithm which works as a finetuning mechanism to enhance rare pragmatic functions values in our training dataset. It allows the model to capture the coexistence of several pragmatic functions while explicitly accounting for their unequal distribution within the same discourse context. Specifically, we approach emoji use not as a simple problem of emoji identification or sentiment detection, but as a multi-label pragmatic classification task in which TBMs learn to associate contextualized textual representations with multiple affective and/or interactional functions expressed through emoji use in ADD. Although MLC has been applied to domains such as “protein function classification, music categorization and semantic scene classification” (Liu et al., 2021, p. 15, see also LeCun et al., 2015; Qasim et al., 2024), CSMLC has received little attention in digital pragmatics, particularly in modeling emoji pragmatic functions in ADD, and this study bridges this gap.

2.2. Positive-class inverse-frequency re-weighting

Label imbalance is commonly addressed through data-level approaches such as over- or under-sampling, and algorithm-level approaches that modify the learning objective (Tyagi & Mittal, 2019). Cost-sensitive learning belongs to the latter category, assigning different costs to classification errors so that errors involving underrepresented classes or labels exert greater influence on model optimization (Rian, 2023). In MLC, this distinction is particularly important as imbalance can vary substantially across labels within the same dataset: a function may be highly frequent, whereas another may occur only rarely, producing different levels of learning difficulty across the output space. This formulation gives larger weights to rare positive labels and smaller weights to frequent ones. It is consistent with the `pos_weight` formulation implemented in binary cross-entropy (BCE) with logits (or `BCEWithLogitsLoss`), where the positive component of the loss is scaled according to the negative-to-positive ratio (cf. Gavrilescu et al., 2025). The principal advantage of this approach is that it changes the contribution of labels to the optimization objective without changing the training data itself. Thus, unlike resampling, it preserves the original instances and their multi-label co-occurrence structure while increasing the learning pressure associated with underrepresented functions. Cost-sensitive re-weighting is consequently well suited to settings in which the objective is to improve minority-label recognition without artificially modifying the empirical distribution (cf. Rian, 2023).

In this study, we conceptualize IFRW as a cost-sensitive learning strategy that adjusts the contribution of positive instances to the multi-label optimization objective according to the empirical frequency of each target function. IFRW does not alter or modify the training data through resampling, it rather preserves the original distribution of pragmatic functions while assigning greater optimization weight to positive labels of underrepresented functions and lower weight to those of highly frequent functions. We further introduce a IFRW algorithm to compute inverse-frequency re-weighting and examine its effectiveness in representing rare micro pragmatic functions. This formulation allows the models to retain the naturally occurring multi-label structure of ADD while explicitly addressing label imbalance during finetuning.

2.3. Arabic BERT-based models

2.3.1. AraBERT

AraBERT is one of the earliest and most influential monolingual pretrained Transformer models developed specifically for Arabic NLP (Antoun et al., 2020). It has been released in several versions (v0.1, v0.2, v1, and v2), with Base and Large configurations designed for different computational requirements and downstream NLP applications (Antoun et al., 2020, 2021). The Base configuration comprises 12 Transformer encoder layers, a hidden size of 768, and 12 self-attention heads, whereas the Large configuration contains 24 encoder layers, a hidden size of 1024, and 16 self-attention heads. AraBERT employs a WordPiece-based tokenizer with an Arabic vocabulary designed to accommodate the morphological and cliticization characteristics of Arabic. The model versions also differ in their preprocessing strategies. Whereas v0.1 and v0.2 are trained on raw, unsegmented Arabic text, v1 and v2 use Farasa-based segmentation before tokenization, separating common prefixes, suffixes, and clitics and thereby helping to address sparsity and tokenization challenges in Arabic (Abdelali et al., 2016; Antoun et al., 2020). During pretraining, AraBERT follows the BERT framework based on masked language modeling and is trained on large-scale Arabic corpora, including news articles, Arabic Wikipedia, OSCAR, and other web-crawled sources amounting to tens of gigabytes (Devlin et al., 2019; Antoun et al., 2020; see also Alkaabi et al., 2025).

In this study, we use AraBERTv2 serving as a representative Arabic TBM, utilizing it to cover part of our diglossic data in SA Facebook posts, and see to what extent a pretrained SA-based TBM can perform on DA-based data in ADD, thus comparing its performance on both overall, multi-label tasks with that of MARBERT, which will be briefed in the section to follow.

2.3.2. MARBERT

MARBERT is a monolingual, bidirectional TBM developed specifically for DA by Abdul-Mageed and Elmadany (2021) to address the linguistic diversity of DA and the unique characteristics of ADD. Unlike Arabert, MARBERT was pretrained on a large-scale corpus of Arabic posts from X, enabling it to capture dialectal variation, colloquial expressions, informal writing styles, and other social media-specific linguistic phenomena. This model consists of 12 Transformer encoder layers and approximately 163 million trainable parameters (Shormani et al., 2026, see also Abdul-Mageed et al., 2022). Owing to its extensive exposure to dialectal and user-generated content, MARBERT has sometimes outperformed several multilingual and Arabic language models, including AraGPT2 (Antoun et al., 2020), CAMELBERT (Abdul-Mageed et al., 2022), and XLM-Roberta (Conneau et al., 2020), on a range of Arabic NLP benchmarks. During pretraining, MARBERT employed the MLM objective on billions of Arabic social media posts, resulting in robust contextual representations that are particularly well suited to downstream tasks such as sentiment analysis, dialect identification, offensive language detection, and social media text classification (Abdul-Mageed et al., 2021; Drous et al., 2025).

In our study, MARBERT, as an optimized TBM for DA/ADD, is compared with AraBERT, which is a SA-oriented model. Comparing its performance with that of AraBERT allows for an empirical assessment of whether

social media-oriented pretraining provides measurable advantages for MLC of emoji pragmatic functions (see also Dainas & Herring, 2017; Yus, 2025).

2.4. Arabic: Diglossia and DL

Arabic, both SA and DA, is known as a rich diglossic language with a rich linguistic system phonologically, morphologically, syntactically, etc. In its diglossic status (see e.g. Ferguson, 1959), Arabic presents substantial computational challenges for NLP due to its complex interaction between morphology, orthography, and diglossia (see e.g., Habash, 2010). Habash (2010) argues that Arabic's rich root-and-pattern morphological system enables the generation of extensive lexical families through the combination of consonantal roots and morphological templates, while the frequent attachment of clitics, including conjunctions, prepositions, articles, pronouns, and particles, allows individual word forms to encode multiple layers of grammatical information. This morphological productivity, although linguistically expressive, creates considerable lexical sparsity and increases the difficulty of tokenization, normalization, and representation learning (Shormani & Alenezi, 2026). These challenges are further intensified in digital communication, where Arabic orthography exhibits substantial variation, including alternative spellings of the same lexical items, inconsistent representation of **harakat** **alif**, **dal**, **ra**, etc. forms, *harakat* **alif**, **dal**, **ra** context-dependent use, omission of diacritics, character elongation like **www**, letter repetition such as **www** (Egyptian lexis for want) for emphasis, and non-standard spellings used to express pragmatic meanings such as emphasis, irony, and emotional intensity. Additionally, diglossia in its 2 forms, H(igh) and L(ow) has shaped ADD, as users frequently move between SA and DA diverse regional dialects, including Egyptian, Gulf, Levantine, Yemeni, Iraqi, and Maghrebi varieties (see also Alhaissoni, et al., 2026). These H and L dialects differ considerably in vocabulary, morphology, syntax, and phonological representation, and social media communication often involves rapid alternation between SA and DA forms within the same interaction (cf. Nashir, et al., 2025).

The diglossic nature of Arabic along with its ADD use perhaps increases its status as a low-resource language, rather difficult for TBMs. It is one of the most actively used languages in digital communication, with millions of users generating vast quantities of digital content daily across social networking platforms such as Facebook, X, Instagram, YouTube, and messaging applications (cf. Mellor, 2022; Hamdan et al., 2021; Shormani & Alenezi, 2026). Differently from well-edited formal, structured data like books, corpora (e.g., falak), these channels' data contain several types of inconsistencies, unstructured data, natural posts. This is perhaps due to the effects of digital discourse, which is characterized by spontaneous, informal, and highly dynamic communication, where users frequently combine textual expressions with emojis, hashtags, hyperlinks, images, abbreviations, and other multimodal resources to convey complex meanings. It then follows that computing these types of data requires models capable of processing not only lexical and syntactic information but also contextual, pragmatic, and culturally embedded communicative signals (Shormani et al., 2026).

Previous studies could be synthesized into 4 themes. The first strand tackles and proposes foundational MLC research, establishing that a single instance may simultaneously receive multiple non-mutually-exclusive labels and developed methods for modeling such label structures. Tsoumakas and Katakis (2007) provided a broad framework for MLC, while Read et al. (2011) introduced classifier chains to exploit

dependencies among labels by transforming MLC into a sequence of binary predictions. A second perspective addresses the problem of imbalanced labels by modifying the learning objective rather than the training data. Bian (2023), for instance, applied cost-sensitive learning in an LSTM-based binary classification task, demonstrating the potential of cost-sensitive objectives for improving prediction under imbalance, whereas Huang et al. (2021) investigated balancing strategies for long-tailed multi-label text classification and highlighted the difficulty of learning rare labels. A third stream has extended MLC and imbalance-aware learning to Arabic NLP. Xu et al. (2025) investigated Arabic multi-label emotion classification using stacked representations from ArabicBERT, MARBERT, and AraBERT together with a hybrid loss incorporating class weighting, label correlations, and contrastive learning. Although its use of Arabic Transformer models and imbalance-aware multi-label learning, the study addresses emotion classification, weighting in their micro statues a study (see also Cui, 2025). Cui (2025) addresses data imbalance in Transformer-based multi-label task, employing sentiment analysis, too, with weighted loss.

The fourth and final juncture is closely related to our study approaching digital pragmatics and computing it through emoji uses. This trend is foreseen by Shormani (2026), which examined whether MARBERT could learn 1 macro Interpersonal pragmatic function, and 5 micro ones, viz., *politeness*, *respect*, *solidarity*, *empathy*, and *encouragement*, from emoji-containing Arabic Facebook posts. This study provides the immediate empirical motivation for this study by demonstrating that MARBERT can learn context-dependent pragmatic functions beyond conventional sentiment or emotion classification. However, it did not formulate emoji pragmatics as our study's CSMLC problem, nor did it examine the effects of severe label imbalance through IFRW or compare AraBERT and MARBERT under controlled baseline and reweighted conditions, which this study handles. However, our study, though building on these studies, among others, adopts the multi-label perspective but differs from these traditional approaches (like those used by Tsoumakas & Katakis, 2007) by using TBMs with independent sigmoid outputs to learn 10 co-occurring pragmatic functions from ADD. Our study also applies IFRW directly within the multi-label loss, specifically to severe imbalance among 10 pragmatic functions while preserving the original multi-label distribution. It also models emoji digital pragmatic functions in their macro and micro hierarchical representations, and employs a controlled AraBERT–MARBERT comparison under the same IFRW procedure.

With this review in mind, 3 major questions are attempted in this study.

1. How much can AraBERT and MARBERT reliably learn pragmatic functions in ADD CSMLC?
2. To what extent can positive-class IFRW improve the identification of underrepresented micro pragmatic functions?
3. Which TBM perform better in this juncture, MARBERT or AraBERT?

3. Methodology

3.1. Data collection and partitioning

We collected a corpus of 11001 Arabic Facebook posts using Python 3. Facebook was selected because it represents one of the most widely used social networking platforms in the Arab world, where users engage in naturally occurring online interactions characterized by informal language, dialectal variation, rich

contextual information, and extensive emoji usage. Facebook posts typically provide sufficient textual context to facilitate reliable interpretation of emoji functions, making the platform particularly suitable for investigating context-dependent pragmatic phenomena (Shormani, 2026).

After collection, the corpus underwent systematic screening and cleaning, though we did not exclude any noise ingredients. Identifying information including usernames, account identifiers, and potentially identifying elements embedded in hashtags or user mentions, was excluded to protect user privacy and comply with relevant platform and research-ethics considerations. After screening and cleaning, the unique corpus consisted of 8390 unique Arabic Facebook posts containing 2 or more emojis. A total of 792 posts without emojis, 1688 posts with 1 emoji, and 131 empty posts containing no meaningful textual content were excluded from our corpus. We removed posts with 1 emoji because they did not satisfy the pragmatic criteria set forth for our study, i.e. they did not include ambiguous textual and/or pragmatic content.

The final corpus was partitioned into 3 mutually exclusive subsets: 5872 training posts, 1260 validation posts, and 1258 evaluation posts. Thus, approximately 70% of the corpus was allocated to training, 15% to validation, and 15% to final evaluation (see also Naghdi & Erfani, 2026). The training set was used for model finetuning and calculation of class-reweighting parameters, the validation set was used for monitoring model performance and model selection during training and the evaluation set was held out exclusively for final performance assessment. Maintaining mutually exclusive subsets prevents information leakage and provides an independent basis for evaluating model generalization (Munaye et al., 2026).

3.2. Annotation and multi-label classification

The corpus was manually annotated according to a pragmatic annotation framework designed specifically to capture the communicative functions performed by emojis in their linguistic and interactional contexts. The framework differs fundamentally from conventional sentiment classification, in which a text is commonly assigned a single dominant emotional category. The annotation scheme thus adopts a multi-label classification framework, allowing multiple labels to be assigned to the same post whenever the contextual evidence supports their simultaneous occurrence. We recruited 3 annotators in the annotation process. They received intensive training for 15 days, with 3 hours of training per day, followed by 41 days of annotation (+ 2 days for resolving disagreement cases). The annotation framework distinguishes between 2 macro pragmatic domains, Affective and Interactional, each comprising 5 micro functions: *prayer*, *sadness*, *surprise*, *anger*, and *love*, *agreement*, *mitigation*, *emphasis*, *humor*, and *sarcasm*. The operational definitions and illustrative emojis are presented in Table 1.

Table 1
An operationalized and annotation framework for coding emoji pragmatics

Macro function	Micro function	Operational definition	Illustrative emojis
Affective	Pray	Expresses e.g., prayer, supplications, gratitude	🙏 🙏 🙏
	Sadness	Expresses e.g., grief, disappointment, or emotional pain	😞 😞 😞
	Surprise	Expresses e.g., astonishment or disbelief	😲 😲 😲
	Anger	Denotes e.g., anger, annoyance, or frustration	😡 😡 😡
	Love	Expresses e.g., warmth, fondness, or emotional closeness	😍 😍 😍
Interactional	Agreement	Indicates e.g., approval or concurrence	👍 👍 👍
	Mitigation	Softens e.g., disagreement, criticism or requests	🙄 🙄 🙄
	Emphasis	Highlights or intensifies a message	🔥 🔥 🔥
	Humor	Creates e.g., amusement or playfulness	😂 😂 😂
	Sarcasm	Signals e.g., ironic or sarcastic intent	🙄 🙄 🙄

It should be noted that the annotation process was hierarchical in nature but did not constitute a hierarchical classification model.

Each post was first assigned 1 macro pragmatic domain, viz., Affective or Interactional to identify its overarching communicative orientation. The post was then annotated for the specific micro pragmatic functions realized by its emojis in context. For supervised learning, each post was represented by a 10-dimensional binary vector, with each dimension corresponding to 1 micro function. A value of 1 indicates that the corresponding pragmatic function is present, whereas 0 indicates its absence. It follows that a post may receive multiple positive labels. For example, a post annotated with *Pray*, *Sadness*, and *Love*, for example, is represented as:

$$[1, 1, 0, 0, 1, 0, 0, 0, 0, 0]$$

Here, the active positions correspond to *pray*, *sadness*, and *love*. The macro category, Affective is not encoded as a separate label because it is already represented by the active micro functions belonging to that category. Similarly, a post annotated with the Interactional functions *humor*, *emphasis*, and *sarcasm* would be represented as:

$$[0, 0, 0, 0, 1, 0, 0, 1, 1, 0]$$

3.3. Class imbalance and cost-sensitive re-weighting

The 10 micro functions exhibit substantial differences in frequency, creating a severe class-imbalance problem. This is particularly important in MLC because standard binary cross-entropy can

disproportionately favor the abundant negative class and frequent positive labels, potentially causing the model to underpredict rare pragmatic functions. To mitigate this problem, we employed cost-sensitive loss re-weighting during finetuning rather than resampling the training data. For each micro function c , a positive-class weight was calculated as the ratio of negative to positive instances in the training set: W_c

$$W_c = \frac{N_{\text{negative},c}}{N_{\text{positive},c}}$$

1

To operationalize Eq. (1), we computed a separate positive-class weight for each of the 10 target micro functions using only the training-set label distribution. The computation procedure is presented in Algorithm 1.

Algorithm 1

Computation of IFRW

Input

Training dataset D with N samples and C target micro-functions; ground-truth binary label matrix $Y \in \{0,1\}^{N \times C}$

Output

Positive-class weight vector $W \in R^C$

1. **1.** Initialize W as an empty vector of size C .
2. **2.** For each micro-function $c = 1, \dots, C$:
3. **3.** $N_{\text{positive},c} = \sum_{i=1}^N Y_{ic}$
4. **4.** $N_{\text{negative},c} = N - N_{\text{positive},c}$
5. **5.** If $N_{\text{positive},c} > 0$:
6. **6.** $W_c = N_{\text{negative},c} / N_{\text{positive},c}$
7. **7.** Else:
8. **8.** $W_c = 1.0$
9. **9.** End if
10. **10.** End for
11. **11.** Return W

where $N_{\text{positive},c}$ denotes the number of training posts annotated with micro function c , and $N_{\text{negative},c}$ denotes the number of training posts not annotated with that function. The resulting weight was incorporated into the BCE loss so that errors involving rare positive labels contributed more strongly to the optimization objective than errors involving highly frequent labels.

This approach was applied identically to both AraBERT and MARBERT. Given that the weights were calculated only from the training-set label distribution, the approach avoids the use of validation or evaluation information during model optimization (cf. Ren et al., 2018). The training-set distribution and corresponding positive-class weights are as shown in Table 2.

Table 2
Training-set distribution and corresponding positive-class weights

Micro function	Positive count	Negative count	Positive weight
Pray	2361	3511	1.4871
Sadness	2539	3333	1.3127
Surprise	79	5793	73.3291
Anger	97	5775	59.5361
Love	2540	3332	1.3118
Agreement	71	5801	81.7042
Mitigation	40	5832	145.8000
Emphasis	3334	2538	0.7612
Humor	3334	2538	0.7612
Sarcasm	3221	2651	0.8230

3.4. Re-weighting of pragmatic functions

As Table 2 displays, the largest weights were assigned to the least frequent micro functions. *Mitigation*, with only 40 positive training instances, received the highest positive-class weight 145.80, followed by Agreement (81.70; 71 positive instances), Surprise (73.33; 79 positive instances), and Anger (59.54; 97 positive instances). These elevated weights increase the contribution of rare positive examples to the loss and reduce the likelihood that underrepresented functions will be ignored during optimization. As for moderate weighting of (relatively) balanced functions, micro functions with comparatively balanced positive and negative distributions required substantially less adjustment. *pray*, with 2361 positive instances, received a weight of 1.49, while *sadness*, with 2539 positive instances, received a weight of 1.31. Similarly, *love* received a weight of 1.31. These values indicate that only moderate compensation was necessary for these relatively frequent functions (cf. Wu et al., 2020; Huang et al., 2021; Yan et al., 2024).

Regarding down-weighting highly frequent functions, the positive classes of highly frequent micro functions were assigned weights below 1. *emphasis* and *humor*, each with 3334 positive instances, received weights of 0.76, while Sarcasm, with 3221 positive instances, received a weight of 0.82. This down-weighting reduces the relative contribution of highly frequent positive labels to the loss and helps prevent the optimization process from being dominated by majority categories (cf. Huang et al., 2021). This re-

weighting strategy provides a cost-sensitive mechanism for balancing the contribution of the 10 pragmatic functions while retaining the original distribution of the training corpus.

3.5. Model optimization and finetuning

Recall that we approached emoji use not as simple sentiment detection, but as a MLC pragmatic task where models learn to associate contextualized text with multiple non-mutually-exclusive Affective and Interactional functions. These were transformed into probabilities, a sigmoid activation function is applied independently to every output neuron (see also Goodfellow et al., 2016; Morozov et al., 2024).

Algorithm 2

Cost-Sensitive Multi-Label Finetuning Procedure

Input

Pre-trained transformer model (AraBERT or MARBERT), training set D_{train} ,
validation set D_{val} , class weight vector W , learning rate η , epochs E

Output

Finetuned model parameters θ^*

1. 1. Initialize model parameters θ
 2. 2. For epoch = 1 to E do:
 3. 3. For each batch (X_{batch} , Y_{batch}) in D_{train} do:
 4. 4. logits = Model(X_{batch} ; θ)
 5. 5. $\hat{y}_{hat}$ = Sigmoid(logits)
 6. 6. // Compute Cost-Sensitive Binary Cross-Entropy Loss using W
 7. 7.
$$\text{Loss} = -\frac{1}{BC} \sum_{i=1}^B \sum_{c=1}^C [W[c] \cdot Y_{batch}[i, c] \log(\hat{y}[i, c]) + (1 - Y_{batch}[i, c]) \log(1 - \hat{y}[i, c])]$$
 8. 8. // Update parameters via AdamW optimizer
 9. 9. θ = OptimizerStep(θ , gradients of Loss, η)
 10. 10. End for
 11. 11. Evaluate model performance on D_{val}
 12. End for
 13. Return best model parameters θ^*
- Our loss function incorporates positive class weights W_c computed inversely proportional to class frequencies ($W_c = N_{negative} / N_{positive}$) as:

$$\text{Loss} = -\frac{1}{BC} \sum_{i=1}^B \sum_{c=1}^C [W_c Y_{i,c} \log(\hat{y}_{i,c}) + (1 - Y_{i,c}) \log(1 - \hat{y}_{i,c})]$$

2

Through this formulation, categories such as *mitigation* or *agreement* were assigned high penalties categories like Emphasis or Humor which are highly frequent were assigned lower penalties. The predicted probabilities converted into binary label assignments with the standard predefined decision threshold $\tau = 0.5$, for each label via inference as (3):

$$\hat{y}_j = \begin{cases} 1 & \text{if } p_j \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

3

AraBERT and MARBERT were finetuned for the MLC pragmatic task. Each model-specific tokenizers were used to tokenize the datasets, with sequences padded/truncated to a maximum length of 128 tokens. The classification layer produced 10 independent output probabilities through sigmoid activation as in (4) (see e.g., Goodfellow et al., 2016; Morozov et al., 2024):

$$\sigma(z_i) = \frac{1}{1 + e^{-z_i}}$$

4

The models were optimized using AdamW and a re-weighted BCE loss appropriate for MLC. Using standardized set of hyperparameters, both TBMs were finetuned for 5 epochs, using a learning rate of 2×10^{-5} and a batch size of 16. All experiments were implemented using Colab and HuggingFace Transformers framework with PyTorch and conducted on an NVIDIA Tesla T4 GPU. AraBERT and MARBERT were trained in separate computational sessions/notebooks to maintain independent model outputs and avoid unintended variable or memory contamination.

4. Results and discussion

4.1. overall performance

We present our findings in this section and discuss them. The average performance on the training workflow summary in Table 3, shows that the baseline models already achieve relatively strong instance-level performance, particularly MARBERT-B, which reaches a Jaccard accuracy of 0.84.

Table 3
Performance on average in training workflow

Model architecture	Jac. Acc.	HamLo	Precision	Recall	F1
AraBERT-B	0.74	0.11	0.46	0.46	0.47
AraBERT-Re-W	0.77	0.10	0.51	0.63	0.53
MARBERT-B	0.84	0.07	0.51	0.52	0.51
MARBERT-Re-W	0.87	0.04	0.72	0.76	0.73

However, their average precision, recall, and F1 remain comparatively modest (0.46–0.52), in their baseline architectures, indicating that strong aggregate overlap does not necessarily correspond to balanced recognition across the 10 pragmatic functions. This pattern is associated with the highly uneven distribution of the target functions, under which frequent categories can dominate aggregate performance while rare categories receive little predictive coverage.

IFRW substantially improves the average performance of both architectures. MARBERT-ReW achieves the highest average precision (0.72), recall (0.76), and F1 (0.73), together with a Jaccard accuracy of 0.87 and Hamming loss of 0.04. The improvement is considerably stronger for MARBERT than AraBERT, suggesting that the effectiveness of cost-sensitive optimization depends partly on the underlying pretrained representation.

As for the test performance, MARBERT-B achieves a Micro-F1 of 0.87 but a substantially lower Macro-F1 of 0.53, while AraBERT-B achieves Micro-F1 of 0.78 and Macro-F1 of 0.49 (Table 4). The large gap between micro- and macro-level performance indicates that the baseline models perform substantially better on frequent functions than on rare ones. Thus, high Micro-F1 should not be interpreted as evidence of uniformly successful multi-label classification. Rather, it can conceal the omission of theoretically important minority functions.

Table 4
Overall Test Metrics

Model architecture	HammLo	Ma-Prec.	Ma-Rec.	Ma-F1	Mi-Prec.	Mi-Rec.	Mi-Fi	Weig-F1	Jac.-Acc.
AraBERT-B	0.13	0.50	0.50	0.49	0.79	0.76	0.78	0.79	0.76
AraBERT-Re-W	0.12	0.52	0.60	0.54	0.82	0.82	0.83	0.82	0.79
MARBERT-B	0.08	0.52	0.53	0.53	0.86	0.87	0.87	0.86	0.85
MARBERT-Re-W	0.06	0.69	0.72	0.75	0.88	0.88	0.88	0.88	0.87

Reweightings changes this distribution of performance, particularly for MARBERT. MARBERT-ReW reaches a Macro-F1 of 0.75, compared with 0.53 for MARBERT-B, while Macro-Precision increases from 0.52 to 0.69 and Macro-Recall from 0.53 to 0.72 (Table 4). At the same time, Micro-F1 remains high at 0.88 and

Hamming loss decreases to 0.06. AraBERT also benefits from reweighting, although to a much smaller extent: Macro-F1 increases from 0.49 to 0.54 and Macro-Recall from 0.50 to 0.60 (cf. Figure 3). These results demonstrate that the principal benefit of reweighting is not simply higher overall prediction, but more equitable recognition across the pragmatic-function inventory. Specifically, the impact of reweighting on Macro-F1 highlights a stark architectural divergence: while AraBERTv2 registers a modest gain from 0.50 to 0.54 (+ 0.04), MARBERT achieves a massive leap from 0.53 to 0.75 (+ 0.18). This substantial difference demonstrates that cost-sensitive optimization yields vastly superior macro-level recovery when paired with a contextual representation better tuned to the heterogeneous register of social media discourse.

4.2. Effect of IFRW

The function-level results provide direct evidence for why the reweighted models achieve higher macro-level performance. The proposed IFRW derives the positive-class weight for each function from the training-set distribution, using the ratio of negative to positive instances. Consequently, rare functions receive substantially larger weights than frequent functions (Table 5).

Table 5
Per-class function performance (micro F1)

Function	AraBERT-B	AraBERT-ReW	MARBERT-B	MARBERT-ReW
Pray	0.75	0.79	0.86	0.83
Sadness	0.76	0.80	0.85	0.86
Surprise	0.00	0.45	0.00	0.73
Anger	0.00	0.07	0.00	0.37
Love	0.77	0.80	0.85	0.87
Agreement	0.00	0.13	0.00	0.56
Mitigation	0.00	0.00	0.00	0.19
Emphasis	0.86	0.83	0.90	0.89
Humor	0.86	0.83	0.90	0.89
Sarcasm	0.84	0.82	0.88	0.88

The resulting function-level changes closely correspond to this design. Eq. (1) changes this optimization pressure in a function-specific manner (Table 2). For example, Mitigation had only 40 positive instances against 5832 negative instances, yielding:

$$W_{Mitigation} = \frac{5832}{40} = 145.80$$

Similarly, *agreement* and *surprise* with 71 positive and 5801 negative, and 79 positive and 5793 negative instances, respectively received:

$$W_{Agreement} = \frac{5801}{71} \approx 81.70$$

$$W_{Surprise} = \frac{5793}{79} \approx 73.33$$

However, *emphasis* and *humor* each had 3334 positive and 2538 negative instances, respectively and thus received:

$$W_{Emphasis} = W_{Humor} = \frac{2538}{3334} \approx 0.76$$

The resulting weights range from 0.76 for *emphasis* and *humor* to 145.80 for *mitigation*, with Agreement, Surprise, and Anger also receiving very high weights (Table 2). This procedure does not modify the original corpus through oversampling or undersampling. It rather changes the relative contribution of positive instances during optimization. Algorithm 1 thus translates the empirical imbalance into a function-specific weight vector, while Algorithm 2 incorporates this vector into the cost-sensitive BCE objective through the positive-class weighting term, whose resulting training and evaluation loss convergence dynamics across epochs are illustrated in Fig. 1.

In this way, the 2 algorithms provide a direct link between the observed distribution of pragmatic functions and the optimization process. The effect is particularly clear for *mitigation*. Only 40 positive instances occur against 5832 negative instances, producing a weight of 145.80. By contrast, *emphasis* and *humor* each contain 3334 positive and 2538 negative instances and thus receive a weight of 0.76. The weighting scheme thus does not simply increase the importance of all positive labels; it redistributes positive-class learning pressure according to empirical rarity.

At the same time, the results show that reweighting cannot fully compensate for an extremely limited number of positive examples. *mitigation* remains difficult for MARBERT-ReW despite receiving the highest weight, achieving an F1 of only 0.19. With only 40 positive training instances, the model has limited contextual evidence from which to learn this function. Thus, cost-sensitive learning can increase the learning emphasis assigned to minority categories, but it cannot substitute entirely for adequate corpus representation.

As illustrated in Fig. 2, reweighting also introduces small, negative F1-score changes (typically between (−0.01 and −0.05)) for several high-frequency categories such as *emphasis*, *humor*, and *pray* in Arabert’s behavior. These minor reductions reflect the optimization trade-off inherent in cost-sensitive reweighting. Thus, as gradient pressure is redistributed toward underrepresented minority functions, the model slightly relaxes its focus on dominant classes. However, these negligible drops on frequent categories are vastly outweighed by the massive macro-level gains achieved across the rare functional inventory.

4.3. MARBERT vs AraBERT

A major finding is the consistent superiority of MARBERT over AraBERT under both training conditions. The baseline difference is relatively modest, with Macro-F1 values of 0.53 and 0.49, respectively. After

reweighting, however, the difference becomes substantially larger, viz., MARBERT-ReW reaches 0.75 compared to 0.54 for AraBERT-ReW (Table 4). The function-level results reveal the same pattern. MARBERT-ReW achieves F1 scores of 0.73 for *surprise*, 0.56 for *agreement*, 0.37 for *anger*, and 0.19 for *mitigation*, compared with 0.45, 0.13, 0.07, and 0.00, respectively, for AraBERT-ReW (Table 5). This pattern suggests an interaction between pretrained representation and cost-sensitive optimization. Reweighting provides stronger optimization pressure for rare functions, but MARBERT appears to provide a more effective contextual representation from which these distinctions can be learned. The relatively modest improvement of AraBERT indicates that reweighting alone cannot fully overcome limitations associated with the underlying representation. A plausible explanation concerns the linguistic characteristics of Arabic social-media discourse. The Facebook corpus may contain SA, DA, informal expressions, and mixed registers. Recall that AraBERT was developed primarily around Arabic resources oriented towards SA, whereas MARBERT was designed with greater exposure to Arabic varieties and user-generated text (Abdul-Mageed & Elmadany, 2021; Antoun et al., 2020). This may make MARBERT better suited to the heterogeneous linguistic environment of social media, where pragmatic interpretation depends on the interaction among lexical choices, discourse context, register, and emoji use.

This explanation should, however, be treated cautiously. The experiments at hand do not directly isolate DA variations or experimentally test the contribution of particular pretraining data. MARBERT’s stronger performance should therefore be interpreted as evidence of greater empirical suitability for our corpus, rather than as definitive evidence that DA TBMs’ pretraining alone caused the observed advantage. Importantly, the findings suggest that the 2 components are complementary rather than interchangeable. MARBERT-B alone failed to recover several rare functions, while reweighting AraBERT produced only modest gains. The strongest results emerged when the stronger contextual representation of MARBERT was combined with function-specific cost-sensitive optimization. Thus, effective Arabic multi-label pragmatic classification appears to require attention to both representation quality and label imbalance.

4.4. Per-function pragmatic performance

The per-function results further clarify the relationship between frequency and model performance (cf. Figure 3).

Both baseline models perform strongly on frequent functions, particularly *emphasis*, *humor*, and *sarcasm*, while producing F1 = 0.00 for *surprise*, *anger*, *agreement*, and *mitigation* (cf. Table 4). This asymmetry demonstrates that the apparent difficulty of some pragmatic functions is closely associated with their limited representation in the training data. Reweighting substantially changes this pattern; MARBERT-ReW recovers performance on all 4 previously unrecognized minority functions, although *mitigation* remains comparatively weak. At the same time, performance on frequent functions declines only slightly. For example, MARBERT’s F1 decreases for *pray* from 0.86 to 0.83, for *emphasis* from 0.90 to 0.89, and *humor* from 0.90 to 0.89. Although the difference is minor, it indicates that reweighting primarily redistributes learning emphasis rather than uniformly increasing or decreasing performance across all functions (cf. also Fig. 2). These findings also reinforce the importance of using macro-averaged evaluation in highly imbalanced multi-label pragmatic classification. A model can obtain a high Micro-F1 by correctly identifying frequent functions while effectively ignoring rare ones. The contrast between MARBERT-B’s Micro-F1 of 0.87

and Macro-F1 of 0.53 illustrates this limitation particularly clearly. Macro-F1 is thus more informative when the objective is to model the full pragmatic repertoire rather than predominantly frequent functions.

To recapitulate, the findings unveil several implications for Arabic NLP including: i) indicating that emoji use can be computationally modeled as a multi-label pragmatic phenomenon, extending Arabic NLP beyond conventional sentiment, emotion, and topic classification towards finer-grained discourse and interpersonal interpretation, ii) demonstrating the importance of evaluating imbalanced Arabic NLP tasks using both macro- and micro-averaged measures. High Micro-F1 alone can conceal complete failure on rare but theoretically meaningful functions. The substantial improvement in Macro-F1 following IFRW thus provides stronger evidence of broad pragmatic function learning than aggregate instance-level performance alone, iii) suggesting that model selection should consider the linguistic ecology of the target corpus. For Arabic social-media data characterized by dialectal variation, informal language, and mixed registers, models with stronger exposure to user-generated Arabic may provide more suitable representations (cf. also Shormani & Alenezi, 2026), and iv) the study demonstrates that representation learning and imbalance-aware optimization are complementary solutions. MARBERT provides a stronger contextual foundation, while IFRW redirects optimization towards underrepresented functions. Their combination produces the strongest overall performance and offers a practical framework for multi-label pragmatic classification in Arabic digital discourse.

5. Conclusions and limitations

To conclude, this study examined the capability of 2 Arabic BERT-based transformer models, namely AraBERT and MARBERT to perform multi-label classification of emoji pragmatic functions within ADD. Leveraging a curated corpus of 8390 unique emoji-containing posts structured into rigorous training, validation, and evaluation splits, the empirical evaluation demonstrates that MARBERT substantially outperformed AraBERT across nearly all global and per-label metrics. This performance advantage is driven by MARBERT's superior alignment with DA as characterized by dialectal, informal, and mixed-register properties of digital discourse (cf. Shormani, 2026). Additionally, the incorporation of inverse-frequency cost-sensitive reweighting effectively addressed severe class imbalance, enabling significant recovery of minority pragmatic functions such as *surprise*, *agreement*, *anger*, and *mitigation* that standard baseline models largely failed to capture.

The findings also reveal a crucial distinction between aggregate and balanced metrics, as high micro-averaged performance consistently exceeded macro-averaged results. This confirms that aggregate metrics are heavily dominated by frequent labels like *emphasis*, *humor*, and *sarcasm*, which can mask underlying deficiencies in rare categories. It follows that achieving a balanced representation across the entire pragmatic inventory involves a deliberate trade-off, where substantial gains in minority-class recall and macro-F1 incur a modest reduction in majority-dominated metrics. These insights yield important practical implications for computational pragmatics, establishing MARBERT as a highly robust backbone for socio-pragmatic analysis and underscoring the necessity of evaluating multi-label models with complementary, label-sensitive metrics under class imbalance.

The study makes 3 main contributions: i) it extends Arabic NLP beyond the predominant emphasis on sentiment, emotion, and polarity classification by formulating emoji use as a MLC pragmatic problem, ii) it applies a positive-class IFRW strategy to address severe imbalance among fine-grained pragmatic functions while retaining the original corpus distribution, and iii) it provides a controlled comparison of AraBERT and MARBERT, offering empirical evidence concerning the relative effectiveness of general Arabic versus dialect- and social-media-oriented pretraining for cost-sensitive multi-label pragmatic classification.

Despite its robust outcomes, this study has several limitations including: i) although cost-sensitive reweighting successfully mitigated skewness, extremely rare pragmatic classes still suffered from limited positive training instances, suggesting that advanced focal loss or synthetic data augmentation could yield further improvements, ii) Arabic diglossic nature, along with DA unconventional spellings, digital discourse inconsistencies, and informal character elongations continue to introduce noise that standard subword tokenizers handle imperfectly, iii) while the evaluated taxonomy captures essential affective and interactional functions, broader socio-pragmatic dimensions such as politeness strategies, face-saving acts, and power dynamics remain ripe for future investigation (see also Shormani, 2026), and iv) this study adopts a flat multi-label framework rather than hierarchical classifiers. While this design successfully isolated and evaluated per-function learning via cost-sensitive reweighting, it does not explicitly model the conditional dependencies between macro- and micro-level pragmatic functions. Future work could explore hierarchical classification architectures or tree-structured loss functions to leverage these structural relationships natively. Future studies could compare IFRW with alternative imbalance-handling approaches, including focal loss, class-balanced loss, adaptive reweighting, and threshold optimization (cf. Ativi, et al., 2026). Such comparisons would establish whether the gains achieved by MARBERT-ReW can be further improved.

Declarations

Availability of data and materials

The training, validation, and evaluation datasets generated or analyzed during the current study will be available on GitHub at [[https://github.com/...BERT-based-models-and cost...](https://github.com/...BERT-based-models-and-cost...)]

Funding

This research project did not receive internal or external funding.

Conflicts of interest

Authors declare no conflicts of interest.

Consent to publish

Authors have approved the final version of the manuscript and agree to its publication.

Ethics statement

This article does not contain any studies with human or animal participants.

Informed consent

There are no human participants in this article and informed consent is not required.

References

1. Abdul-Mageed M, Elmadany A (2021) ARBERT & MARBERT: Deep bidirectional transformers for Arabic. *Proceedings of ACL-IJCNLP (Long Papers)*, 7088–7105. <http://dx.doi.org/10.18653/v1/2021.acl-long.551>
2. Abdul-Mageed M, Elmadany M, Nagoudi EMB (2022) CAMELBERT: A Collection of Pretrained Models for Arabic NLP. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 628–640)
3. Alhaissoni AS, Yafooz W, Alsaeedi A (2026) Automatic speech recognition and dialect identification for Arabic and Saudi dialects: a systematic literature review. *J King Saud Univ Comput Inform Sci* 38(5):190. <http://dx.doi.org/10.1007/s44443-026-00609-6>
4. Alkaabi HAA, Jasim kadhim, A., Darroudi A (2025) Arabic NLP: A Survey of Pre-Processing and Representation Techniques. *J Comput Sci Inform Technol Telecommunication Eng* 6(2):876–890
5. Annepaka Y, Pakray P (2025) Large language models: a survey of their development, capabilities, and applications. *Knowl Inf Syst* 67(3):2967–3022. <http://dx.doi.org/10.1007/s10115-024-02310-4>
6. Antoun W, Baly F, Hajj H (2020) AraBERT: Transformer-based model for Arabic language understanding. *Proceedings of the 4th workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection*, 9–15. <https://doi.org/10.48550/arxiv.2003.00104>
7. Antoun W, Baly F, Hajj H (2021) AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the sixth Arabic natural language processing workshop* (pp. 196–207)
8. Ativi WX, Chen W, Kwao L, Ayivi W, Sam F, Alqahtani A, Al-Antari MA (2026) CMAF-Net: cross-modal attention fusion with information-theoretic regularization for imbalanced breast cancer histopathology. *Sci Rep* 16(1):9607. <https://doi.org/10.1038/s41598-025-32794-1>
9. Bian Y (2023) A LSTM and Cost-sensitive learning-based real-time warning for civil aviation over-limit. In *2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE)* (pp. 550–556). IEEE
10. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Stoyanov V (2020), July Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). <http://dx.doi.org/10.18653/v1/2020.acl-main.747>
11. Cui X (2025) Addressing data imbalance in transformer-based multi-label emotion detection with Weighted Loss. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)* (pp. 247–256)

12. Dainas AR, Herring SC (2021) Interpreting emoji pragmatics. In *Pragmatics & Beyond New Series* (pp. 107–144). John Benjamins Publishing Company. <https://doi.org/10.1075/pbns.318.04dai>
13. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
14. Drous HA, Barakat H, Atta AB (2025) Finetuning MARBERT for sentiment analysis in Jordanian Arabic dialects using a synthetic dialectal corpus. *J Comput Sci Technol Stud* 7(11):373–386. <http://dx.doi.org/10.32996/jcsts.2025.7.11.36>
15. Ferguson CA (1959) Diglossia word 15(2):325–340. <http://dx.doi.org/10.1093/oso/9780195092905.003.0002>
16. Gavrilescu M, Leon F, Minea AA (2025) Techniques for transversal skill classification and relevant keyword extraction from job advertisements. *Information* 16(3):167. <http://dx.doi.org/10.3390/info16030167>
17. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press
18. Habash NY (2010) Introduction to Arabic natural language processing. Morgan & Claypool. <http://dx.doi.org/10.1007/978-3-031-02139-8>
19. Hamdan HJ, Al-Salman S (2021) The use of Arabic neologisms in social media applications. *Int J Arabic-English Stud* 21(1). <http://dx.doi.org/10.33806/ijaes2000.21.1.3>
20. Hou W, Wang J, Lin Q, Wang X, Huang L (2025) Improving clinical foundation models with multi-modal learning and domain adaptation for chronic disease prediction. *IEEE J biomedical health Inf* <http://dx.doi.org/10.1109/JBHI.2025.3595140>
21. Huang Y, Giledereli B, Köksal A, Özgür A, Ozkirimli E (2021) Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on empirical methods in natural language processing* (pp. 8153–8161). <http://dx.doi.org/10.18653/v1/2021.emnlp-main.643>
22. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444. <http://dx.doi.org/10.1038/nature14539>
23. Lin N, Qin G, Wang G, Zhou D, Yang A (2023) An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 8730–8744). <http://dx.doi.org/10.1038/nature14539>
24. Liu W, Wang H, Shen X, Tsang IW (2021) The emerging trends of multi-label learning. *IEEE Trans Pattern Anal Mach Intell* 44(11):7955–7974. <http://dx.doi.org/10.1109/TPAMI.2021.3119334>
25. Mellor N (2022) Arab digital journalism. Taylor & Francis. <http://dx.doi.org/10.1109/TPAMI.2021.3119334>
26. Morozov AV, Levkivskyi VL, Plechystyy DD (2024) Activation functions in neural networks: Overview and comparison. *Sci Notes VI Vernadsky Taurida Natl Univ Series: Tech Sci* 35(74). <http://dx.doi.org/10.32782/2663-5941/2024.3.1/22>

27. Mostafavi S, Yahyavi Y, Ravanmehr R (2026) Systematic literature review on sentiment analysis using transformers. *Int J Data Sci Analytics* 22(1):49
28. Munaye YY, Gebeyehu AD, Tai LC, Abebe ZA, Workneh AB, Tarekegn RB, Tarekegn GB (2026) Machine learning-driven intrusion detection for securing IoT-based wireless sensor networks. *Future Internet* 18(2):113. <http://dx.doi.org/10.1007/s41060-026-01021-z>
29. Naghdi M, Erfani A (2026) Multimodal and temporal modeling of highway–rail grade crossing crash severity using explainable artificial intelligence. *Reliab Eng Syst Saf* 113081. <http://dx.doi.org/10.1016/j.ress.2026.113081>
30. Narayan S (1997) The generalized sigmoid activation function: Competitive supervised learning. *Inf Sci* 99(1–2):69–82. [http://dx.doi.org/10.1016/S0020-0255\(96\)00200-9](http://dx.doi.org/10.1016/S0020-0255(96)00200-9)
31. Nashir WA, Mohsen AM, Al-Shargabi AA, Nour MK, Al-Onazi BB (2025) Noor: A pipeline framework for Classical Arabic parsing. *IEEE Access*. <http://dx.doi.org/10.1109/ACCESS.2025.3607915>
32. Qasim AN, Alani S, Mahmood SN, Mohammed SS, Aziz DA, Ata KIM (2024) Enhancing brain stroke detection: A novel deep neural network with weighted binary cross entropy training. *Revue d'Intelligence Artificielle* 38(3):777. <http://dx.doi.org/10.18280/ria.380304>
33. Rahali A, Akhloufi MA (2023) End-to-end transformer-based models in textual-based NLP. *Ai* 4(1):54–110. <http://dx.doi.org/10.3390/ai4010004>
34. Read J, Pfahringer B, Holmes G, Frank E (2011) Classifier chains for multi-label classification. *Mach Learn* 85(3):333–359. <https://doi.org/10.1007/s10994-011-5256-5>
35. Ren M, Zeng W, Yang B, Urtasun R (2018) Learning to reweight examples for robust deep learning. In *International conference on machine learning* (pp. 4334–4343). PMLR
36. Shormani MQ (2025) What fifty-one years of linguistics and artificial intelligence research tell us about their correlation: A scientometric analysis. *Artif Intell Rev* 58(12):379. <http://dx.doi.org/10.1007/s10462-025-11332-5>
37. Shormani MQ (2026) Does Machine know interpersonal pragmatics? Evidence from MARBERT's learning of emoji pragmatics in Arabic digital discourse. *arXiv preprint arXiv:2608.01174*. <https://doi.org/10.48550/arXiv.2608.01174>
38. Shormani MQ, Hassan IA, Alshawsh MY (2026) Machine learning and emoji prediction: How much accuracy can MARBERT achieve? *arXiv preprint arXiv:2604.21108*. <https://doi.org/10.48550/arXiv.2604.21108>
39. Shormani MQ, Alenezi MA (2026) Arabic Nominalization as Form of Pragmatic Actions in Digital Discourse: A Corpus-Based Study. *Corpus Pragmat* 10(1):35. <http://dx.doi.org/10.1007/s41701-026-00237-5>
40. Tsoumakas G, Katakis I (2007) Multi-label classification: An overview. *Int J Data Warehouse Min* 3(3):1–13. <https://doi.org/10.4018/jdwm.2007070101>
41. Tyagi S, Mittal S (2019) Sampling approaches for imbalanced data classification problem in machine learning. In *Proceedings of ICRIC 2019: Recent innovations in computing* (pp. 209–221). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-030-29407-6_17

42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008
43. Wang J, Dai X, Shi B, Liu M, Peng Y (2026) Topic-aware transformer with hierarchical prompting learning for multi-label image classification. PeerJ Comput Sci 12:e3552. <http://dx.doi.org/10.7717/peerj-cs.3552>
44. Xu Y, Aslam MA, Jun W, Ahmed N, Zaman MI, Hamza M, Aslam S (2025) Improving Arabic multi-label emotion classification using stacked embeddings and hybrid loss function. IEEE Access 13:188939–188954. <http://dx.doi.org/10.1109/ACCESS.2025.3580478>
45. Yus F (2025) Emoji pragmatics. Springer Nature
46. Zhang M-L, Zhou Z-H (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837. <https://doi.org/10.1109/TKDE.2013.39>

Figures

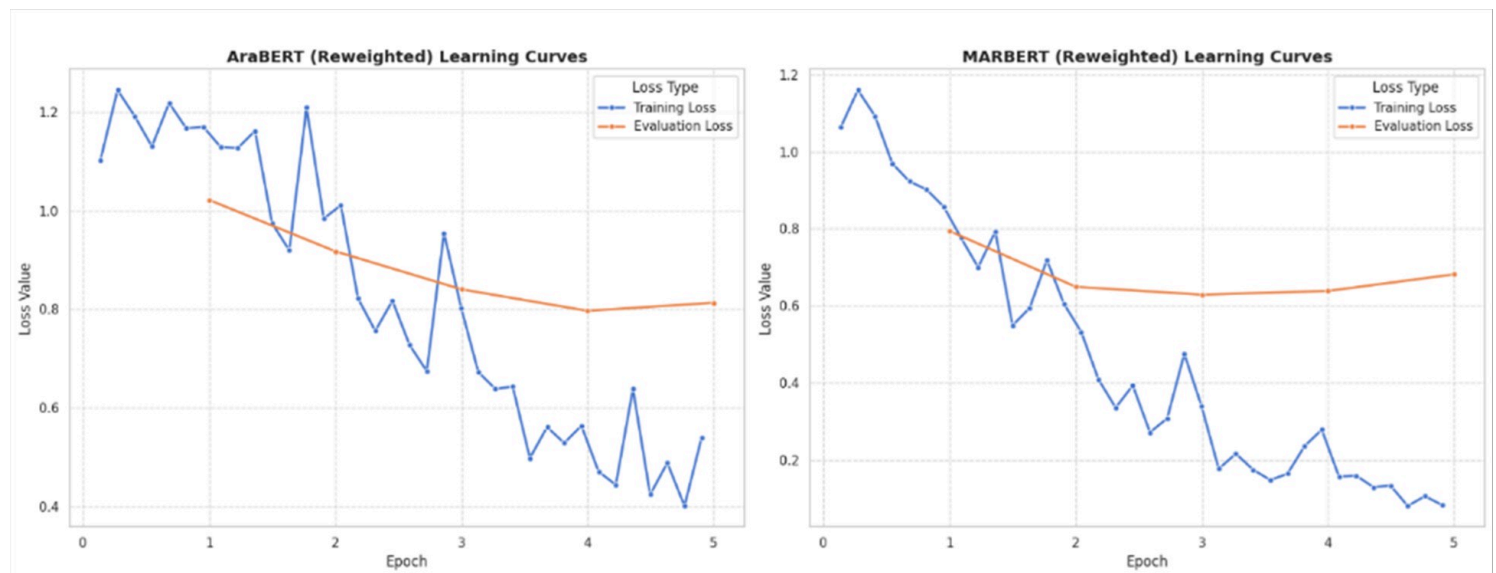


Figure 1

Learning and validation losses

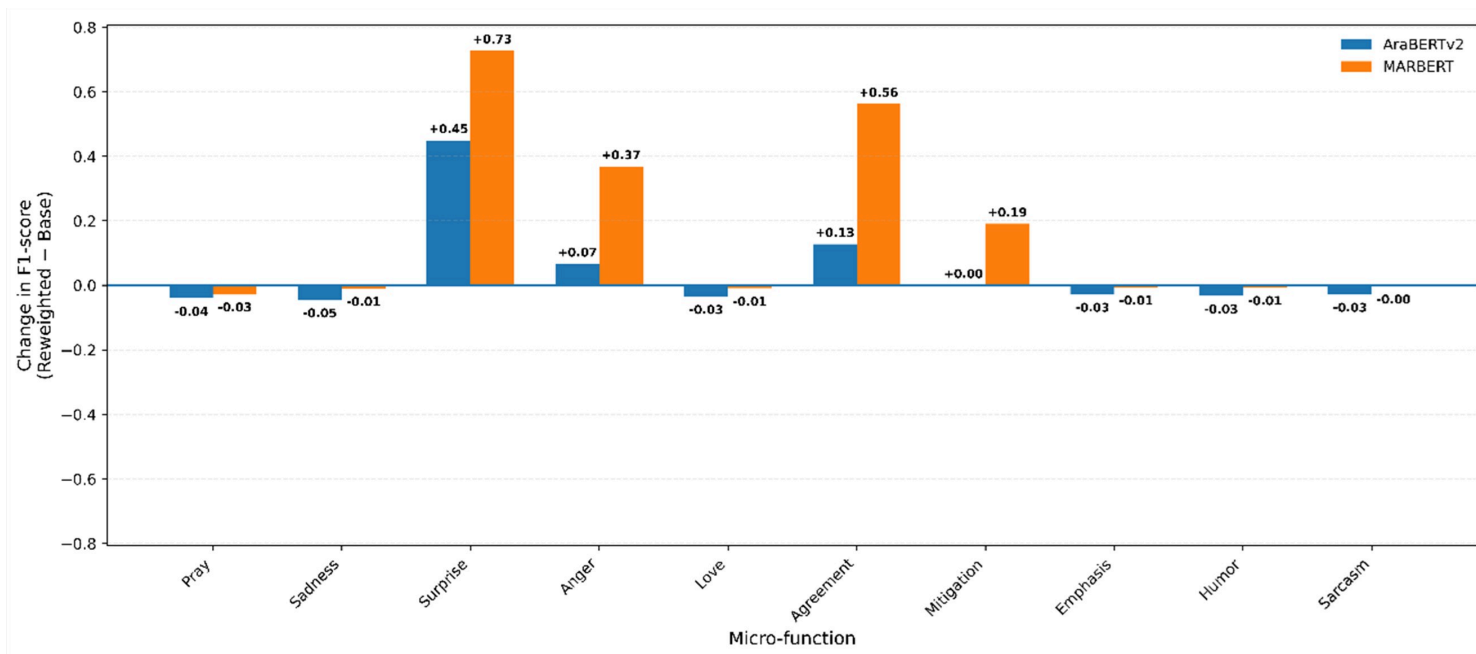


Figure 2

Effect of IFRW

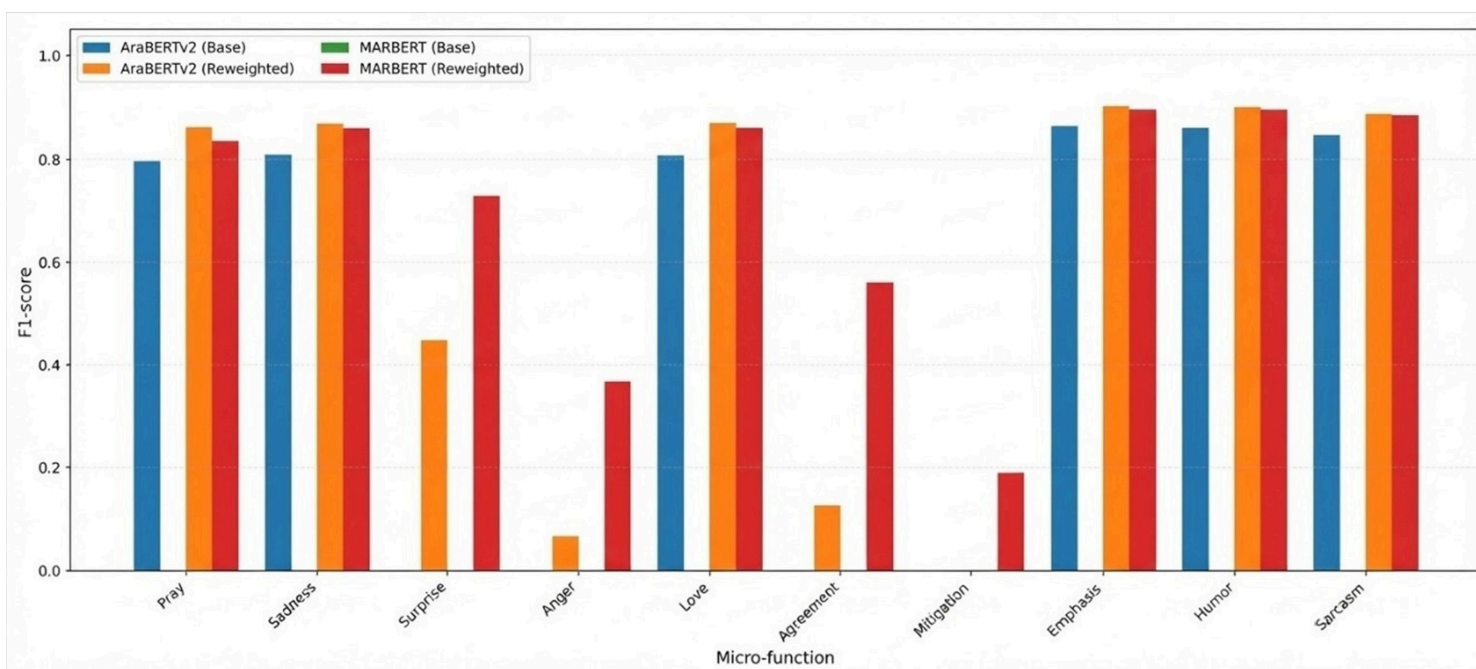


Figure 3

Per-function pragmatic performance