A Multi-Task Learning Approach to Dialectal Arabic Identification and Translation to Modern Standard Arabic

Abdullah Khered[1,2], **Youcef Benkhedda**[1] **and Riza Batista-Navarro**[1]
[1]The University of Manchester, UK
[2]King Abdulaziz University, Saudi Arabia
abdullah.khered@manchester.ac.uk
youcef.benkhedda@manchester.ac.uk
riza.batista@manchester.ac.uk

Abstract

Translating Dialectal Arabic (DA) into Modern Standard Arabic (MSA) is a complex task due to the linguistic diversity and informal nature of dialects, particularly in social media texts. To improve translation quality, we propose a Multi-Task Learning (MTL) framework that combines DA-MSA translation as the primary task and dialect identification as an auxiliary task. Additionally, we introduce LahjaTube, a new corpus containing DA transcripts and corresponding MSA and English translations, covering four major Arabic dialects: Egyptian (EGY), Gulf (GLF), Levantine (LEV), and Maghrebi (MGR), collected from YouTube. We evaluate AraT5 and AraBART on the Dial2MSA-Verified dataset under Single-Task Learning (STL) and MTL setups. Our results show that adopting the MTL framework and incorporating LahjaTube into the training data improve the translation performance, leading to a BLEU score improvement of 2.65 points over baseline models.

1 Introduction

Machine Translation (MT) is a Natural Language Processing (NLP) task that aims to translate between natural languages automatically. Over the last decade, Neural Machine Translation (NMT) has improved translation quality by leveraging deep learning to model complex linguistic patterns from large datasets. A widely used NMT architecture is the Sequence-to-Sequence (Seq2Seq) model, which consists of an encoder-decoder framework typically based on Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU)(Cho et al., 2014). The encoder processes the input sentence into a compressed representation, which the decoder then uses to generate the translated output(Sutskever et al., 2014). More recently, Transformer-based models have surpassed earlier Seq2Seq architectures by replacing recurrence with self-attention and parallel computation (Vaswani et al., 2017), resulting in faster translation, improved accuracy, and better handling of long-range dependencies. Furthermore, pre-trained Transformer-based models have demonstrated state-of-the-art performance across various NLP tasks beyond machine translation, solidifying the Transformer as the dominant architecture in modern NLP research (Qiu et al., 2020).

Despite these advancements, low-resource language translation remains a challenge, particularly for Dialectal Arabic (DA) to Modern Standard Arabic (MSA) translation. Arabic operates in a diglossic environment: MSA is the standardised form used in education, media, and formal communication, while DA is the informal variant shaped by regional cultures, local expressions, and daily communication (Salloum et al., 2014; Sadat et al., 2014). The challenge in DA-MSA translation lies in the variability across Arabic dialects. Each dialect has morphological and syntactic differences, often incorporating borrowed words from other languages and region-specific expressions (Mallek et al., 2017). Moreover, the rise of social media has further complicated these challenges, as Arabic speakers frequently mix dialects, use slang, abbreviations, emojis, and code-switching with other languages (Alruily, 2020).

To address these challenges, fine-tuning Arabic pre-trained Transformer models such as AraBART (Kamal Eddine et al., 2022) and AraT5 (Elmadany et al., 2023) on dialect-specific corpora has proven beneficial in overcoming data scarcity for DA-MSA translation (Khered et al., 2025). Moreover, Multi-Task Learning (MTL) has emerged as a promising approach for enhancing DA-MSA translation. Instead of training a model solely for translation, MTL enables joint training on multiple related

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 21–31, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-004

tasks, such as MSA-English translation (Baniata et al., 2018b), Part-Of-Speech (POS) tagging (Baniata et al., 2018a) and translation of multiple dialects (Moukafih et al., 2021). These auxiliary tasks provide additional linguistic signals that help improve model generalisation, contextual understanding, and robustness to informal variations. Our research builds upon normalising Arabic text in social media and improving the results on the Dial2MSA-Verified dataset (Khered et al., 2025) by integrating an MTL framework. We also introduce the LahjaTube dataset, a new corpus sourced from YouTube videos, to enrich model training. LahjaTube was developed to address the shortage of DA-MSA translation datasets, particularly those that include informal and real-world language from major Arabic dialects as commonly found on social media. Our objectives include:

- Develop an MTL framework for DA-MSA translation that incorporates dialect identification as an auxiliary task.
- Automatically collect and construct the LahjaTube dataset, covering four major Arabic dialects: Egyptian (EGY), Gulf (GLF), Levantine (LEV), and Maghrebi (MGR) with their corresponding MSA and English translations.
- Evaluate the performance of MTL models on both DA-MSA translation and dialect identification tasks using the Dial2MSA-Verified dataset. This includes training on different combinations of datasets, incorporating the newly introduced LahjaTube corpus.

The code for the MTL framework and supplementary material for this paper are available online at `https://github.com/khered20/MTL-Dial2MSA`.

2 Related Work

MTL is a machine learning technique that jointly trains multiple tasks, allowing knowledge sharing between related tasks (Zhang and Yang, 2017). MTL has been explored in various NLP tasks (Kumar et al., 2019; Chen et al., 2024) including those with limited data resources (Mamta et al., 2022; Guzman et al., 2024; Elgamal et al., 2024), leading to more generalised representations.

In Arabic, MTL has been applied to various linguistic tasks, including diacritic restoration, where auxiliary tasks such as word segmentation and POS tagging have been utilised to enhance accuracy (Alqahtani et al., 2020). Similarly, dialect identification has benefited from MTL approaches, with hierarchical attention mechanisms improving fine-grained classification at the city, state, and country levels (Abdul-Mageed et al., 2019). Moreover, MTL has been integrated with pre-trained language models such as MARBERT (Abdul-Mageed et al., 2021) for Arabic dialect identification at both the country and province levels, demonstrating that sharing task-specific attention layers improves generalisation across Arabic varieties (El Mekki et al., 2021). Additionally, Arabic Natural Language Understanding (ANLU) has been enhanced through MTL frameworks that facilitate parameter sharing across multiple tasks. This approach has led to a notable performance on some tasks within the ALUE benchmark, highlighting the importance of carefully considering task relationships and loss scaling (Alkhathlan and Alomar, 2024).

Recent studies in MTL with MT have revealed that incorporating auxiliary tasks can improve translation performance (Zaremoodi et al., 2018; Pham et al., 2023). In the context of DA-MSA translation, various MTL approaches have been proposed. Baniata et al. (2018b) explored a unified multitask NMT model where DA-MSA translation served as the main task and MSA-English translation as the auxiliary task. The architecture utilised a separate encoder for each task whilst sharing a single decoder. Another study by Baniata et al. (2018a) further improved MTL for Arabic dialect translation by integrating POS tagging as an auxiliary task. This model adopted a shared-private Bi-LSTM-CRF architecture, encoding DA sentences and segment-level POS tags. The results demonstrated that the POS tagging task improved the translation BLEU score. Similarly, Moukafih et al. (2021) adopted a seq2seq MTL framework, encoding and decoding pairs of different dialects and MSA within the PADIC-parallel dataset (Meftouh et al., 2018) using a shared GRU model. Their Many-to-One sitting improved the translation performance, surpassing statistical MT models in 88% of translation cases.

Our research focuses on Arabic social media normalisation, specifically translating DA into MSA within the Dial2MSA-Verified dataset (Khered et al., 2025). The Dial2MSA-Verified, an extension of Dial2MSA (Mubarak, 2018), is a multi-reference dataset covering tweets from four di-

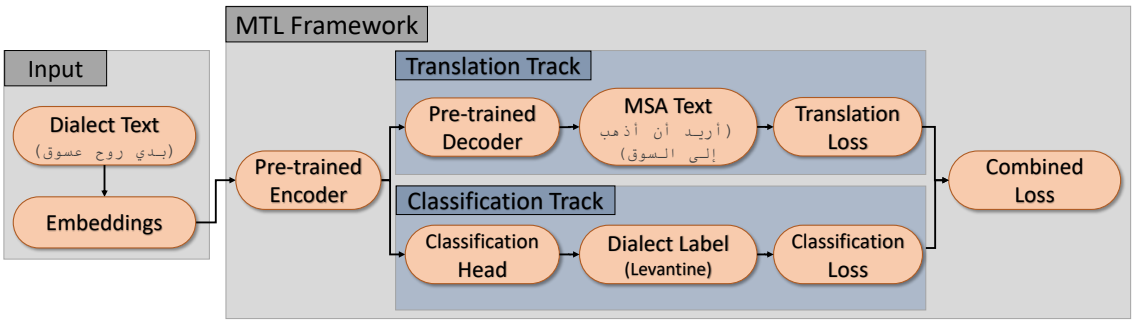

Figure 1: Architecture of the MTL framework processes a shared pre-trained encoder, followed by two parallel tasks: a translation track that generates MSA text using a pre-trained decoder and a classification track that predicts the dialect label

alects: EGY, MGR, GLF and LEV dialects with their multiple MSA translations. Khered et al. (2025) further explores joint and independent training strategies, demonstrating that joint training across dialects leads to superior translation performance. Additionally, transformer-based models, including AraT5 (Elmadany et al., 2023) and AraBART (Kamal Eddine et al., 2022), have been benchmarked, with AraT5 emerging as the best-performing model. In this context, we propose a novel MTL framework that leverages Arabic-specific pre-trained Transformer models (AraT5, AraBART) for DA-MSA translation and dialect identification. Furthermore, we introduce LahjaTube, a dataset containing four Arabic dialects, each with multiple transcripts, along with their corresponding MSA and English translations to enrich the training data.

3 MTL for DA-MSA Translation and Dialect Identification

In this section, we present our Multi-Task Learning (MTL) framework designed to simultaneously perform two tasks: DA-MSA translation and dialect identification, as illustrated in Figure 1. This framework is built upon state-of-the-art Transformer models and optimises both tasks through shared representations. The dataset used in this study consists of Arabic dialectal sentences paired with their corresponding MSA translations and dialect labels. The labels encompass the four dialects: EGY, GLF, LEV, MGR as well as MSA.

3.1 Architecture Overview

The architecture is based on a Transformer encoder-decoder model that is specifically pre-trained in Arabic Language, such as AraT5 and AraBART, enhanced with an additional classification head. The architecture consists of the following components:

- **Encoder** converts the input DA text into a high-dimensional vector representation, which is shared between the translation and classification tasks.
- **Decoder** generates the corresponding MSA translation from the encoder's representation.
- **Classification Head** is added to the encoder output to perform dialect classification.

3.2 Loss Functions

Our model is trained using an MTL approach that combines two objectives: dialect classification and Seq2Seq translation. To achieve this, we define two separate loss functions and combine them into a weighted objective function.

Classification Loss For dialect classification, we use a separate classification head, which applies a linear transformation followed by a softmax activation. The model predicts the probability distribution over C dialect classes. The classification loss is formulated using Cross-Entropy Loss:

$$\mathcal{L}_{\text{classification}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \quad (1)$$

C is the total number of classes (four dialects and MSA). y_c is the true label, represented as a one-hot vector. $\hat{y}_c$ is the predicted probability for class c.

Translation Loss For the translation task, we also use the Cross-Entropy Loss, which measures the difference between the predicted probability distribution and the actual target sequence. The translation loss is defined as:

$$\mathcal{L}_{\text{translation}} = -\sum_{t=1}^{T}\sum_{v=1}^{V} y_{t,v}\log(\hat{y}_{t,v}) \quad (2)$$

T is the target sequence length and V is the vocabulary size. $y_{t,v}$ is a one-hot vector representing the true token at position t. $\hat{y}_{t,v}$ is the predicted probability for token v at position t.

Combined Loss Function To train the model jointly for both tasks, we define a weighted combination of the translation and classification losses:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{translation}} + (1-\alpha)\mathcal{L}_{\text{classification}} \quad (3)$$

where α is a hyperparameter in the range $(0, 1)$ that controls the relative importance of the two tasks. This combined loss ensures the model learns the classification and translation tasks simultaneously.

4 LahjaTube Dataset

In this section, we introduce the LahjaTube dataset, a collection of transcripts from YouTube videos covering DA from four Arabic-speaking regions: EGY, GLF, LEV, and MGR. These transcripts are accompanied by English translations which were translated into MSA. The LahjaTube dataset is available upon request for academic purposes.

4.1 Data Collection

The data collection concentrated on YouTube videos created by content creators who speak one of the four aforementioned dialects. We employed the YouTube Data API v3[1] to select videos from countries representative of these dialects. The filtering functions were used to include only videos under Creative Commons Attribution licenses and contain subtitles from both Arabic and English. Once an initial set of videos was identified, we explored the creators' other videos that met our specific criteria. We collected a total of 1,912 videos distributed across the four selected dialects. For the caption extraction process, we used the YouTube Video Subtitles Scraper from the Apify platform[2] to retrieve both the original Arabic transcripts and their corresponding English translations.

4.2 Data Processing and Cleaning

To ensure the quality of the extracted data, we undertook several cleaning and preprocessing steps. First, each sample was defined according to the timestamped segments provided by YouTube subtitles, so that each instance in our dataset corresponds to an English subtitle segment as determined by the video's original caption timing. If subtitles occurred with minimal time gaps and without sentence-final punctuation, we merged them into a single sample. In cases where a single subtitle segment contained multiple short sentences separated by in-line punctuation, we kept these grouped as a single data instance. We also removed any subtitle containing fewer than four words (in either the dialectal Arabic or English lines) to reduce potential noise and ensure sufficient linguistic content. Furthermore, the geographic location of a video's creator alone does not guarantee the actual dialect of the transcripts, as the creator could use different dialects or MSA, host guests from other regions, or produce videos while travelling. We addressed this by applying a dialect identification model to verify the dialect used in each line, ensuring our dataset includes only transcripts where the identified dialect corresponds with the creator's known dialect. The model used is MTL-AraBART, the high-performing dialect identification model produced in this study, trained on the same datasets used in Khered et al. (2025).

4.3 MSA Translation

To enable translation from DA to MSA, we generated MSA translations by translating the English subtitles into MSA using the few-shot GPT-4o[3] model via its API. For each dialect, we designed a specific prompt that included three few-shot examples, which were manually selected from our collected DA–English subtitle pairs. Native Arabic speakers provided accurate MSA translations for these selected dialectal samples, and these few examples were incorporated into the GPT-4o prompt. As illustrated in Figure 2, [Dialect] specifies the relevant dialect, while [DA] and [EN] refer to the original DA transcript and its English translation, respectively.

4.4 Corpus Statistics

The final corpus comprises a total of 31,938 transcripts from YouTube videos distributed across the four aforementioned dialects, along with their English and MSA translations. Subsequently, these transcripts are distributed as shown in Table 1, capturing a variety of dialect-specific expressions and

[1] https://developers.google.com/youtube/v3

[2] https://apify.com/

[3] https://platform.openai.com/docs/models/#gpt-4o

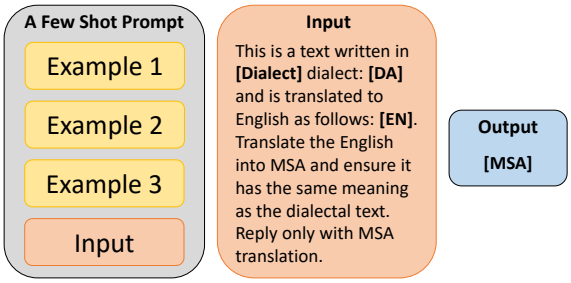

Figure 2: Few-shot prompting strategy used to convert English translations into MSA using the GPT-4o model

vocabulary. Table 1 provides detailed statistics including the size of each dialect corpus, along with the total and unique word counts for both the DA and the corresponding MSA translations.

Dialect	Size	Total Words	Unique Words	Total MSA Words	Unique MSA Words
EGY	10,279	110,387	21,417	112,470	21,613
GLF	7,762	106,669	13,269	112,277	16,192
LEV	7,695	98,138	15,996	102,010	16,851
MGR	6,202	95,148	16,329	95,829	17,439

Table 1: Statistics for LahjaTube corpus

Table 2 highlights several examples from the LahjaTube dataset. In some cases, such as the LEV example, the text might start or end suddenly because it could be part of a larger conversation. Despite this, the English and MSA translations accurately capture the original meaning of the DA conversation, ensuring that all versions convey the same text.

EGY	وبالتالي ممكن انك تشيل الطاقة وتعوض مكانها بالكلام ده
MSA	وبالتالي يمكنك أن تزيل الطاقة وتستبدلها بهذا الكلام.
EN	Therefore, you can remove the energy and replace it with this talk.
GLF	والله لاصدمه خله يكلمني والله لاصدمه طيب قل له خمسين ريال
MSA	والله لأفاجئه، دعه يكلمني، والله لأفاجئه. حسناً، قل له خمسين ريال
EN	I swear I will shock him. Let him talk to me. I swear I will shock him. Okay, tell him fifty riyals.
LEV	معك من القران بس ما بيعرف شو الحكم الفقهي بهال
MSA	معك من القرآن لكنه لا يعرف ما هو الحكم الفقهي في هذا
EN	You know from the Qur'an, but he does not know what the jurisprudential ruling is on this
MGR	المنافع ديالو ذاكشي علاش جربتو ولله الحمد لقيت عليه نتيجه دابا
MSA	فوائده هي السبب الذي جعلني أجربه، ولله الحمد وجدت نتائج الآن
EN	Its benefits are why I tried it, and thank God I found results now.

Table 2: DA transcripts with their MSA and English translations from LahjaTube dataset

4.5 Human Evaluation of MSA Translations

We conducted a human evaluation on a subset of 200 DA-MSA translation pairs from LahjaTube, with 50 samples per dialect. For each dialect, one annotator, a native speaker of the relevant dialect, evaluated only samples from their own dialect. The evaluation followed the multi-dimensional method proposed by Sadiq (2025), which assessed accuracy, fluency, style and tone, cultural suitability, and terminology on a 1-5 scale. As shown in Table 3, the MSA translations in LahjaTube showed high overall quality, with average ratings above 4.5 across most criteria and dialects.

Dialect	Acc	Flu	S&T	Cult	Term	Average
EGY	4.42	4.3	3.7	4.32	4.22	4.19
GLF	4.74	4.72	4.12	4.82	4.74	4.63
LEV	4.62	4.6	4.12	4.74	4.82	4.58
MGR	4.62	4.6	4.16	4.72	4.74	4.57

Table 3: Average human evaluation scores (Acc = Accuracy, Flu = Fluency, S&T = Style & Tone, Cult = Cultural Suitability, Term = Terminology) for DA-MSA translation on a LahjaTube subset (N=50 per dialect, 200 samples)

5 Experimental Design

We conduct experiments using the MTL structure, where DA-MSA translation forms the primary task, and dialect identification serves as an auxiliary task. This structure allows the model to leverage information about the dialect during the translation process, potentially improving translation accuracy.

5.1 Dataset

The Dial2MSA-Verified dataset (Khered et al., 2025) is a multi-reference evaluation dataset, fully verified and sourced from social media, specifically built for DA-MSA translation. Additionally, we integrate the newly introduced LahjaTube dataset, which was created from YouTube video transcripts based on the same four dialects. For all experiments reported in this work, we evaluated our models on the same fixed development and test sets from Dial2MSA-Verified. As summarised in Table 4, the test set contains 2,000 samples per dialect, with some dialect sentences paired with two or three MSA translation references. To assess the impact of the diversity in training data on model performance, we experimented with three different training subsets as presented in Table 4:

- **Subset 1**: The same training set used in Khered et al. (2025), which serves as a baseline. It includes Dial2MSA-Verified-train along with the following additional resources: PADIC (Meftouh et al., 2018), MADAR-train (Bouamor et al., 2018), Arabic STS (Al Sulaiman et al., 2022), and Emi-NADI (Khered et al., 2023) datasets. This is to compare the performance of our new MTL models against previous models.
- **Subset 2**: This training set combines the Dial2MSA-Verified-train and LahjaTube datasets. The goal of this subset is to evaluate the effectiveness of our newly introduced LahjaTube corpus for DA-MSA translation.
- **Subset 3**: The comprehensive training set, incorporating LahjaTube with all training data from Subset 1. This setup aims to produce the most effective translation models by leveraging the largest dataset available.

	Dataset	EGY	GLF	LEV	MGR
Subset 1	Dial2MSA-V-train	9,099	6,575	4,101	3,312
	PADIC	0	0	12,824	25,648
	MADAR-train	13,800	15,400	18,600	29,200
	Arabic STS	2,758	2,758	0	0
	Emi-NADI	0	2,712	0	0
	Total-train-1	**25,657**	**27,445**	**35,525**	**58,160**
Subset 2	Dial2MSA-V-train	9,099	6,575	4,101	3,312
	LahjaTube	10,279	7,762	7,695	6,202
	Total-train-2	**19,378**	**14,337**	**11,796**	**9,514**
Subset 3	Training Subset 1	25,657	27,445	35,525	58,160
	LahjaTube	10,279	7,762	7,695	6,202
	Total-train-3	**35,936**	**35,207**	**43,220**	**64,362**
	Dial2MSA-V-dev	200	200	200	200
	Dial2MSA-V-test	2000 3-R	2000 3-R	2000 2-R	2000 2-R

Table 4: Dataset setup showing the sizes of the three training subsets. In the test set, R indicates the number of reference MSA translations per DA sentence

5.2 Model Configurations and Training Setup

In this study, we used the second version of **AraT5**[4] model (Elmadany et al., 2023), which is based on the T5 architecture (Raffel et al., 2020). AraT5 has a 12-layer encoder and decoder with 768 hidden units per layer. We also used **AraBART** (Kamal Eddine et al., 2022) model, based on the BART architecture (Lewis et al., 2020), which features a 6-layer encoder and a 6-layer decoder, each with 768 hidden units. Both models are pre-trained on large-scale Arabic corpora and further modified by adding a classification head to the encoder's output for our multi-task setup. The additional classification head enables dialect classification, and its loss is calculated separately, as detailed in Section 3. We generated additional pairs from the MSA targets by replicating them as both source and target sentences. These newly created pairs were assigned to the MSA class. Each model is trained under two different settings:

- **Single-Task Learning (STL)**: The model is trained exclusively for DA-MSA translation, serving as the baseline.
- **Multi-Task Learning (MTL)**: The model is trained jointly for DA-MSA translation and dialect identification.

5.3 Evaluation Metrics

We evaluate model performance using both translation and dialect classification metrics. For translation, we use the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and chrF++ (Popović, 2017) scores, both implemented in SacreBLEU (Post, 2018). For dialect classification, we report accuracy, which measures the percentage of correctly predicted dialects, Macro-F1, which computes the F1-score for each class and averages them equally, and Weighted-F1, which adjusts for class imbalance by weighting each class's F1-score based on the number of true instances.

5.4 Hyperparameter Optimisation

Each experiment is conducted under the same hyperparameter settings to ensure fair comparisons. The configurations include a batch size of 16, a learning rate of 5e-5, a maximum sequence length of 128, and training for up to 20 epochs, with early stopping applied if the best BLEU score on the validation set does not improve for three consecutive epochs. BLEU was chosen as the primary metric for early stopping since this study focuses on translation quality. All experiments are run on two Nvidia V100 GPUs. For the MTL setup, the combined loss function, introduced in Section 3, is optimised using weighting values of 0.3, 0.5, and 0.8 to examine the effect of different weighting schemes.

6 Results and Discussion

In this section, we analyse the results of our experiments for DA-MSA translation, comparing the performance of the baseline STL models (STL-AraT5

[4] https://huggingface.co/UBC-NLP/AraT5v2-base-1024

and STL-AraBART) from Khered et al. (2025) with our proposed MTL models. For the dialect identification task, we use as a baseline the results reported by Khered et al. (2025) for an ensemble of multiple fine-tuned MARBERT models (Abdul-Mageed et al., 2021). The MARBERT ensemble was trained and optimised using the hyperparameter described by Khered et al. (2022). We compare these results against our proposed MTL models. While we experimented with different values of α, all results reported in this section are based on the combined loss with $\alpha = 0.5$, which achieved stable performance on both tasks on the development set.

6.1 DA-MSA Translation

Table 5 measures the translation performance using BLEU and chrF++ scores across the three proposed training subsets. For Training Subset 1, the MTL models generally outperform STL models from Khered et al. (2025) in terms of both BLEU and chrF++ scores, particularly for the GLF, LEV and MGR dialects. The overall average BLEU score for MTL-AraT5 reaches 42.23, compared to 41.12 for STL-AraT5, while chrF++ increases from 62.05 to 62.84, highlighting the benefits of incorporating dialect classification as an auxiliary task.

	Model		EGY	GLF	LEV	MGR	Avg
Training Subset 1	STL-AraT5	BLEU	**30.94**	53.96	45.37	34.24	41.12
		chrF++	**52.94**	70.86	65.40	58.99	62.05
	STL-AraBART	BLEU	29.87	51.38	43.07	32.95	39.32
		chrF++	52.26	69.49	64.13	58.12	61.00
	MTL-AraT5	BLEU	29.71	**55.31**	**48.39**	**35.53**	**42.23**
		chrF++	52.21	**72.01**	**67.19**	**59.95**	**62.84**
	MTL-AraBART	BLEU	29.52	53.79	45.98	33.12	40.60
		chrF++	52.31	70.96	66.08	58.48	61.96
Training Subset 2	STL-AraT5	BLEU	29.96	54.21	46.99	34.50	41.42
		chrF++	52.34	71.43	66.52	59.44	62.71
	STL-AraBART	BLEU	26.25	50.44	43.18	32.93	38.20
		chrF++	49.11	69.15	63.94	58.11	60.08
	MTL-AraT5	BLEU	**30.70**	**55.94**	**48.42**	**35.77**	**42.71**
		chrF++	**52.90**	**72.35**	**67.47**	**60.36**	**63.27**
	MTL-AraBART	BLEU	28.41	51.15	43.56	32.88	39.00
		chrF++	50.98	69.28	64.24	57.80	60.58
Training Subset 3	STL-AraT5	BLEU	31.00	54.27	47.68	35.16	42.03
		chrF++	53.19	71.54	66.86	59.78	62.84
	STL-AraBART	BLEU	29.96	53.37	46.64	32.70	40.67
		chrF++	52.34	70.70	66.19	57.57	61.70
	MTL-AraT5	BLEU	**31.73**	**56.51**	**50.31**	**36.55**	**43.77**
		chrF++	**53.81**	**72.71**	**68.54**	**60.77**	**63.96**
	MTL-AraBART	BLEU	29.70	54.41	46.81	34.18	41.27
		chrF++	52.49	70.77	66.19	58.66	62.03

Table 5: The translation performance of STL vs. MTL models evaluated on the Dial2MSA-Verified test dataset, where the results of STL models on Training Subset 1 are from Khered et al. (2025)

In Training Subset 2, which includes only samples from the Dial2MSA-Verified training set and the new LahjaTube dataset, MTL models continue to outperform STL models, with MTL-AraT5 achieving the highest BLEU score of 42.71 and a chrF++ score of 63.27. Surprisingly, this model outperforms models trained on the larger Training Subset 1, highlighting the effectiveness of the LahjaTube dataset in improving DA-MSA translation.

The highest performance is observed in Training Subset 3, where MTL-AraT5, fine-tuned on all training datasets, achieves the best overall results, with a BLEU score of 43.77 and a chrF++ score of 63.96. This demonstrates that combining diverse datasets further improves translation quality.

6.2 Dialect Identification

Table 6 presents the classification performance of the ensemble MARBERT baseline, as reported in Khered et al. (2025), alongside the results of our proposed MTL models (MTL-AraT5 and MTL-AraBART) on the Dial2MSA-Verified test dataset. While MTL-AraBART consistently achieves the highest overall performance, a drop in the Macro-Average F1-score is observed in all MTL models compared to MARBERT. This drop is likely due to the training setup: MARBERT was trained exclusively on the four dialect classes, whereas the MTL models were trained on both the four dialects and MSA. Despite this, MTL-AraBART achieves the best results on other metrics, with an accuracy of 98.85% and a Weighted-F1 score of 99.10%, all obtained with Training Subset 2, which includes only the Dial2MSA-Verified-train and LahjaTube datasets. These results highlight that dialect identification also benefits from the MTL framework.

	Model	Acc	M-F1	W-F1
Training Subset 1	MARBERT	96.950	**96.942**	96.942
	MTL-AraT5	95.750	77.867	97.334
	MTL-AraBART	**97.275**	78.447	**98.059**
Training Subset 2	MTL-AraT5	98.213	78.943	98.678
	MTL-AraBART	**98.850**	**79.284**	**99.104**
Training Subset 3	MTL-AraT5	98.450	79.125	98.906
	MTL-AraBART	**98.688**	**79.273**	**99.091**

Table 6: Dialect identification performance of the ensemble MARBERT baseline and our proposed MTL models (MTL-AraT5 and MTL-AraBART), evaluated on the Dial2MSA-Verified test dataset using Accuracy (Acc), Macro-Average F1 (M-F1), and Weighted-Average F1 (W-F1) scores

6.3 Model Impact on Translation Quality

The results highlight the advantages of the MTL approach for DA-MSA translation, demonstrating consistent improvements over STL models. Translation performance improved when the weighting

parameter prioritised DA-MSA translation while still incorporating dialect identification (e.g., α = 0.5 or 0.8). In contrast, setting α to 0.3 resulted in a decline in performance, likely due to the classification task receiving greater importance, reducing the model's focus on translation. Among the architectures, MTL-AraT5 emerges as the most effective for DA-MSA translation, likely due to AraT5's pretraining on a more extensive and diverse Arabic dataset. Additionally, the results highlight the significance of the training dataset size and diversity, as larger and more varied training datasets enhance translation performance.

6.4 Error Analysis

To evaluate the impact of MTL on DA-MSA translation, we conducted a comparative analysis between the STL-AraT5 model and its multitask-enhanced version, MTL-AraT5. Both models usually produce similar translations, often differing by only one or two words. In many cases, these words had multiple valid translations, making it difficult to determine a single correct output. MTL-AraT5 consistently provides more contextually appropriate translations, likely due to the additional integration of dialect classification, which enhances the model's ability to differentiate and preserve dialect-specific meanings. However, misclassification occasionally affected translation performance. For example, when the model misclassified the input as MSA, it assumed no translation was needed, leading to the reproduction of the original dialectal sentence instead of converting it to MSA.

Despite MTL-AraT5 demonstrating improvements in handling idiomatic expressions and dialect-specific phrases, STL-AraT5 performed better in some instances, particularly in more straightforward lexical mappings. BLEU score comparisons reinforce these findings, indicating that while both models achieve comparable overall performance, MTL-AraT5 excels in dialect-sensitive contexts, whereas STL-AraT5 sometimes provides more direct and literal translations.

7 Conclusion and Future Work

This paper proposed an MTL framework for DA-MSA translation, integrating dialect identification as an auxiliary task. To support this research, we introduced LahjaTube, a new dataset of YouTube video transcripts covering four major Arabic dialects with their corresponding MSA and English translations. Our experiments with AraT5 and AraBART showed that MTL improves translation performance, particularly when LahjaTube is included in the training. MTL-AraT5 achieves the best overall translation performance, outperforming both STL models and MTL-AraBART, with a BLEU score of 43.77 and a chrF++ score of 63.96 when trained on the most comprehensive dataset (Training Subset 3). Meanwhile, MTL-AraBART consistently achieved the highest performance in dialect classification, reaching 98.85% accuracy and a weighted-F1 score of 99.10% in Training Subset 2. These results indicate that both tasks, DA-MSA translation and dialect identification, benefit from the MTL approach, as incorporating dialect identification helps improve translation quality while translation modelling enhances dialect classification. Despite these improvements, challenges remain in handling transliterated words, informal expressions, and code-switching. Additionally, optimising the balance between translation and classification tasks is an area for further research.

Building on our findings, future research can explore several directions to enhance DA-MSA translation. Expanding training data with additional dialectal resources and data augmentation techniques, can improve generalisation. Additionally, utilising large language models (LLMs) with decoder-only Transformer architecture, such as LLaMA, Gemma, and Jais, could improve DA-MSA translation by taking advantage of their strong language understanding and transfer learning abilities.

Limitations

Despite the promising results of our MTL framework, several limitations remain. Although LahjaTube introduces a new source of dialectal data, its coverage may be uneven, potentially underrepresenting certain countries within each dialectal region. While GPT-4o was used to generate MSA translations from English, most translations have not undergone manual verification, and only a small subset was reviewed through human evaluation; thus, some errors or inconsistencies may remain in the automatic MSA translations, which could reduce overall quality. Furthermore, although integrating dialect identification as an auxiliary task improves translation performance, misclassifying DA sentences as MSA can lead to incorrect outputs, with the model simply reproducing the input instead of providing a proper translation.

Ethical Considerations

The LahjaTube dataset consists of transcriptions from publicly available YouTube videos. To ensure ethical and legal compliance, we exclusively collected content licensed under Creative Commons, which permits reuse, including speech transcription for research purposes. Furthermore, we verified that the dataset does not include personal, sensitive, or harmful content. Moreover, the MSA translations were generated automatically from the English transcripts using the GPT-4o model. No manual correction was performed on the entire dataset; however, to assess the translation quality and support the reliability of LahjaTube, we conducted a human evaluation of a small subset of 200 DA-MSA translation pairs.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim A. Elmadany, Arun Rajendran, and Lyle H. Ungar. 2019. Dianet: BERT and hierarchical attention multi-task learning of fine-grained dialect. *CoRR*, abs/1910.14243.

Mansour Al Sulaiman, Abdullah M. Moussa, Sherif Abdou, Hebah Elgibreen, Mohammed Faisal, and Mohsen Rashwan. 2022. Semantic textual similarity for modern standard and dialectal arabic using transfer learning. *PLOS ONE*, 17(8):1–14.

Ali Alkhathlan and Khalid Alomar. 2024. Armt-tnn: Enhancing natural language understanding performance through hard parameter multitask learning in arabic. *Int. J. Know.-Based Intell. Eng. Syst.*, 28(3):483495.

Sawsan Alqahtani, Ajay Mishra, and Mona Diab. 2020. A multitask learning approach for diacritic restoration. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8238–8247, Online. Association for Computational Linguistics.

Meshrif Alruily. 2020. Issues of dialectal saudi twitter corpus. *The International Arab Journal of Information Technology*, 17:367–374.

Laith Baniata, Seyoung Park, and Seong-Bae Park. 2018a. A multitask-based neural machine translation model with part-of-speech tags integration for arabic dialects. *Applied Sciences*, 8:2502.

Laith Baniata, Seyoung Park, and Seong-Bae Park. 2018b. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational Intelligence and Neuroscience*, 2018.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The madar arabic dialect corpus and lexicon. In *International Conference on Language Resources and Evaluation*.

Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multitask learning in natural language processing: An overview. *ACM Comput. Surv.*, 56(12).

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Abdellah El Mekki, Abdelkader El Mahdaouy, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. BERT-based multi-task model for country and province level MSA and dialectal Arabic identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 271–275, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Salman Elgamal, Ossama Obeid, Mhd Kabbani, Go Inoue, and Nizar Habash. 2024. Arabic diacritics in the wild: Exploiting opportunities for improved diacritization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14815–14829, Bangkok, Thailand. Association for Computational Linguistics.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Octopus: A multitask model and toolkit for Arabic natural language generation. In *Proceedings of ArabicNLP 2023*, pages 232–243, Singapore (Hybrid). Association for Computational Linguistics.

Erick Mendez Guzman, Viktor Schlegel, and Riza Batista-Navarro. 2024. Towards explainable multi-label text classification: A multi-task rationalisation framework for identifying indicators of forced labour. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 98–112, Miami, Florida, USA. Association for Computational Linguistics.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. AraBART: a pretrained Arabic sequence-to-sequence model for abstractive summarization. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 31–42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abdullah Khered, Ingy Abdelhalim, Nadine Abdelhalim, Ahmed Soliman, and Riza Theresa Batista-Navarro. 2023. Unimanc at nadi 2023 shared task: A comparison of various t5-based models for translating arabic dialectical text to modern standard arabic. In *Proceedings of ArabicNLP 2023*, pages 658–664.

Abdullah Khered, Ingy Yasser Hassan Abdou Abdelhalim, and Riza Batista-Navarro. 2022. Building an ensemble of transformer models for Arabic dialect classification and sentiment analysis. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 479–484, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abdullah Khered, Youcef Benkhedda, and Riza Batista-Navarro. 2025. Dial2MSA-verified: A multi-dialect Arabic social media dataset for neural machine translation to Modern Standard Arabic. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 50–62, Abu Dhabi, UAE. Association for Computational Linguistics.

Abhishek Kumar, Asif Ekbal, Daisuke Kawahra, and Sadao Kurohashi. 2019. Emotion helps sentiment: A multi-task model for sentiment and emotion analysis. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Fatma Mallek, Billal Belainine, and Fatiha Sadat. 2017. Arabic social media analysis and translation. *Procedia Computer Science*, 117:298–303. Arabic Computational Linguistics.

Mamta, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Exploring multi-lingual, multi-task, and adversarial learning for low-resource sentiment analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Karima. Meftouh, Salima Harrat, and Kamel Smaïli. 2018. PADIC: extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, Antalya, Turkey.

Youness Moukafih, Nada Sbihi, Mounir Ghogho, and Kamel Smaili. 2021. Improving machine translation ofnbsp;arabic dialects through multi-task learning. In *AIxIA 2021 Advances in Artificial Intelligence: 20th International Conference of the Italian Association for Artificial Intelligence, Virtual Event, December 13, 2021, Revised Selected Papers*, page 580590, Berlin, Heidelberg. Springer-Verlag.

Hamdy Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. In *OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 49–53.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hai Pham, Young Jin Kim, Subhabrata Mukherjee, David P. Woodruff, Barnab,s Pczos, and Hany Hassan Awadalla. 2023. Task-based moe for multitask multilingual machine translation. *CoRR*, abs/2308.15772.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63:1872–1897.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of arabic dialects in social media. SoMeRA '14, page 3540, New York, NY, USA. Association for Computing Machinery.

Saudi Sadiq. 2025. Evaluating english-arabic translation: Human translators vs. google translate and chatgpt. *Journal of Languages and Translation*, 12(1):67–95.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 772–778, Baltimore, Maryland. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661, Melbourne, Australia. Association for Computational Linguistics.

Yu Zhang and Qiang Yang. 2017. An overview of multi-task learning. *National Science Review*, 5(1):30–43.