Dial2MSA-Verified: A Multi-Dialect Arabic Social Media Dataset for Neural Machine 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

Social media has become an essential focus for Natural Language Processing (NLP) research due to its widespread use and unique linguistic characteristics. Normalising social media content, especially for morphologically rich languages like Arabic, remains a complex task due to limited parallel corpora. Arabic encompasses Modern Standard Arabic (MSA) and various regional dialects, collectively termed Dialectal Arabic (DA), which complicates NLP efforts due to their informal nature and variability. This paper presents Dial2MSA-Verified, an extension of the Dial2MSA dataset that includes verified translations for Gulf, Egyptian, Levantine, and Maghrebi dialects. We evaluate the performance of Seq2Seq models on this dataset, highlighting the effectiveness of state-of-the-art models in translating local Arabic dialects. We also provide insights through error analysis and outline future directions for enhancing Seq2Seq models and dataset development. The Dial2MSA-Verified dataset is publicly available to support further research [1].

1 Introduction

The rapid growth in social media users has established it as an area of interest for Natural Language Processing (NLP) research. Normalising social media texts' content is transforming informal text into a more standardised form that aligns with established linguistic conventions. This process is a challenging NLP task for morphologically rich languages such as Arabic, especially when parallel corpora for Arabic social media and their corresponding standard forms are limited (Mubarak, 2018).

Arabic, a widely spoken global language, exists in two primary forms: Modern Standard Arabic (MSA) and various regional dialects, collectively known as Dialectal Arabic (DA). MSA, the standardised form of the Arabic language, is utilised in formal contexts such as education, media, literature, and official documentation. As a linguistic bridge across the Arab world, MSA promotes a shared understanding and cultural cohesion among diverse Arab communities. In terms of grammar and vocabulary, MSA follows strict standardised rules, ensuring consistency in formal communication. Conversely, DA is the language of daily interaction, prevalent in informal settings and deeply reflective of the cultural and social identities unique to each region and community (Sadat et al., 2014).

The significant variation between Arabic dialects further complicates NLP tasks, as models trained on MSA alone may struggle with the language used on social media. Arabic users on these platforms tend to use their local informal dialect. A single Arabic word may indicate different interpretations based on the context of the sentence, between two dialects or between a dialect and MSA (Mallek et al., 2017), which shows why it is important to normalise text used in social media. Such a text often combines MSA, dialects, non-Arabic words, and unconventional spelling and may include slang, abbreviations, shortened or compound words, perhaps with grammar or spelling mistakes (Alruily, 2020). Figure 1 demonstrates the issues in Arabic social media text and their correct format.

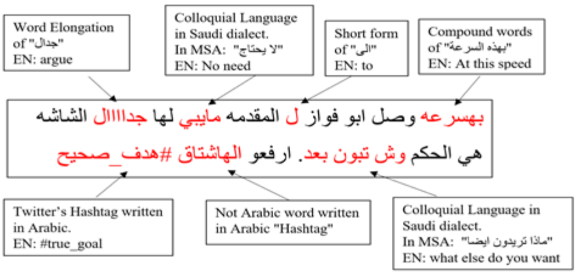

Word Elongation of "جدال" EN: argue
Colloquial Language in Saudi dialect. In MSA: "لا يحتاج" EN: No need
Short form of "الى" EN: to
Compound words of "بهذه السرعة" EN: At this speed
بهسرعه وصل ابو فواز ل المقدمه مايبي لها جدااال الشاشه
هي الحكم وش تبون بعد. ارفعو الهاشتاق #هدف_صحيح
Twitter's Hashtag written in Arabic. EN: #true_goal
Not Arabic word written in Arabic "Hashtag"
Colloquial Language in Saudi dialect. In MSA: "ماذا تريدون ايضا" EN: what else do you want

Figure 1: Examples of issues in Arabic social media text and their correct MSA/normalised forms

[1] https://github.com/khered20/Dial2MSA-Verified

Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4), pages 50–62
January 20th, 2025. ©2024 Association for Computational Linguistics

The increase of unstructured text from various sources, including social media, has highlighted the need for effective preprocessing and normalisation to enhance data quality and usability. While basic preprocessing methods can still handle some issues in Figure 1, others, such as dialectal variations, syntax mistakes, ambiguity, and polysemy, require advanced techniques. To address these issues and the normalisation task, we adopted a Sequence to Sequence (Seq2Seq) technique, specifically using Neural Machine Translation (NMT) architectures. Seq2Seq models have shown promising results in handling translations across different language pairs, including from and to MSA. However, NMT models require large amounts of parallel data (i.e., pairs of sentences in two languages) for effective training. The limited availability of DA-MSA datasets poses an obstacle (Slim and Melouah, 2024). Moreover, in the context of social media, the Dial2MSA dataset (Mubarak, 2018) is currently the only publicly available resource that covers multiple regional dialects, and it has not been fully verified.

In this paper, we present Dial2MSA-Verified, which is built upon Dial2MSA (Mubarak, 2018), a Seq2Seq dataset from social media that encompasses four dialects: Egyptian (EGY), Maghrebi (MGR), Levantine (LEV), and Gulf (GLF). Our contributions to this dataset are two-fold:

- Verifying the dialects that were not verified in the original Dial2MSA dataset, specifically the LEV and GLF dialects, using three human annotators for each. The final dataset was separated into 18,991 tweets for training, 800 for validation, and 8,000 for testing, with multiple MSA references for each tweet.

- Testing and reporting the performance of different Seq2Seq translation models on each dialect of Dial2MSA-Verified, with models such as AraT5v2 (Elmadany et al., 2023) performing particularly well on the GLF dialect and slightly less effectively on other dialects.

2 Related Works

2.1 Seq2Seq DA to MSA translation

Machine Translation (MT) technology has seen significant advancements in recent years, with various approaches and techniques developed across different domains. While existing MT systems supporting Arabic have achieved moderate success, there is a growing focus on improving translation quality and developing more effective technologies, particularly through the application of NMT methods (Zakraoui et al., 2021; Bensalah et al., 2021). For DA translation, two main areas were investigated: DA-English and DA-MSA (Harrat et al., 2019). Multiple works on DA-English translation used MSA as a pivoting between DA and English to address the Out-Of-Vocabulary (OOV) issue in Arabic dialects and to improve the translation (Sawaf, 2010; Salloum and Habash, 2013; Sajjad et al., 2013; Salloum and Habash, 2014; Aminian et al., 2014). Additionally, Salloum et al. (2014) used dialect identification for MT system selection, with MSA as a pivot, to optimise translation between DA and English.

For DA-MSA, early research used rule-based MT (Al-Gaphari and Al-Yadoumi, 2010; Salloum and Habash, 2012; Hamdi et al., 2013), Statistical Machine Translation (SMT) (Salloum and Habash, 2011; Ghoneim and Diab, 2013; Meftouh et al., 2018) and hybrid approaches (Tachicart and Bouzoubaa, 2014). Later systems adapted NMT by either translating one dialect to MSA or multiple dialects to MSA. In single-DA to MSA translation, Al-Ibrahim and Duwairi (2020) employed an RNN Seq2Seq encoder-decoder model to translate the Jordanian dialect into MSA. Slim et al. (2022) applied a transductive Transfer Learning (TL) approach for translating the Algerian dialect to MSA using seq2seq models. Faheem et al. (2024) combined supervised and unsupervised NMT methods to enhance the translation from the EGY dialect to MSA. In multi-DA translation, Shapiro and Duh (2019) conducted training on transformer-based models across different Arabic varieties, including EGY and LEV dialects and MSA. Their findings indicated that leveraging multi-DA datasets can improve the translation quality for other unencountered dialects. Additionally, Baniata et al. (2021) investigated the translation between multiple dialects and MSA by employing a word-piece model to generate sub-word units for input features in the NMT transformer model.

Recently, three shared-tasks were created for DA-MSA translation: the fourth and fifth NADI shared-tasks (Abdul-Mageed et al., 2023, 2024) and OSACT DA-MSA MT shared-task (Elneima et al., 2024). Participants were allowed to use any available dataset and encouraged to create new datasets to train their models. As a result, some teams used a Large Language Model (LLM) such as ChatGPT from OpenAI to augment the training dataset (Khered et al., 2023; AlMusallam and Ahmad, 2024). Participants experimented with various NMT models, such as fine-tuning transformer-based pre-trained in Arabic models.

2.2 Arabic Social Media Normalisation

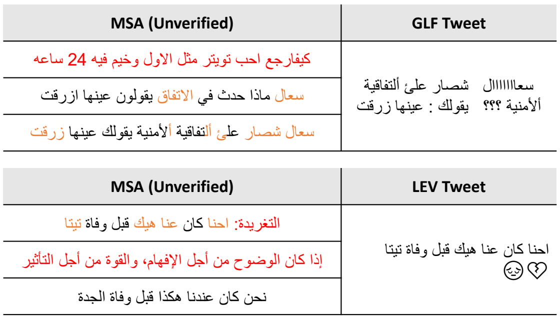
The social media normalisation task involves standardising various linguistic expressions in social media content. This task has attracted research attention across numerous languages and domains (ERYİĞİT and TORUNOĞLU-SELAMET, 2017; Zarnoufi et al., 2020; Aliero et al., 2023). However, these approaches cannot be applied directly to other languages or domains due to linguistic diversity (Matos Veliz et al., 2021). For Arabic, several works have tackled the issue of unstructured text in social media as part of addressing other NLP tasks. For instance, in Sentiment Analysis (SA), Rizkallah et al. (2018) translated some Saudi dialect vocabularies into MSA using the Social Analytics dynamic-link library (DLL) from "AlKhawarizmy Software" and Hegazi et al. (2021) focused on providing a single framework to handle different issues related to preprocessing Arabic tweets. Some studies used the MSA as a pivot language between the DA-English translation in social media. For example, Mallek et al. (2017) used a dictionary of non-standard words and their corresponding MSA to reduce the OOV issue in Arabic tweets, which were then translated into English using a SMT approach. Other studies were focused on normalising single DA on social media, such as Duwairi (2015), which constructed a lexicon for Jordanian DA words and their corresponding MSA. Hamada and Marzouk (2018) created a hybrid system to translate EGY to MSA in social media as part of the ALMoFseH project. They combined naive Bayesian learning to disambiguate morphological analysis, a rule-based transfer mechanism, and a dictionary look-up system. Chennafi et al. (2022) conducted experiments on various tasks within Aspect-Based SA, incorporating a Seq2Seq model for normalisation. The Seq2Seq normalisation model was trained on subsets from the PADIC (Meftouh et al., 2018) and MADAR (Bouamor et al., 2018) datasets to address the OOV issue in EGY sentences.

The Arabic social media normalisation task in previous works was concentrated on a single DA. They applied traditional MT methods to enhance the accuracy of other NLP tasks without being evaluated. Furthermore, the limited use of NMT methods is due to the lack of Seq2Seq data availability from social media platforms. In our research, we proposed a Dial2MSA-Verified, an evaluation dataset of multiple Arabic dialects in social media, by completing the verification of the Dial2MSA dataset. We experimented with various transformer-based NMT models to be evaluated on the Dial2MSA-Verified dataset.

3 Datasets

3.1 Dial2MSA Dataset

The Dial2MSA dataset comprises MSA translations of tweets from four Arabic dialects. The dataset was constructed by initially collecting 175 million Arabic tweets, from which 24,000 tweets were selected based on dialect-specific keywords: 6,000 each for EGY, MGR, LEV, and GLF dialects. The dataset's development involved two annotation tasks: first, human translators provided multiple MSA versions for each tweet; second, these translations underwent verification to remove inaccurate translations and retain only the correct ones. While all four dialects were subjected to the initial translation process, the verification step was completed only for the EGY and MGR dialects, leaving the MSA translations of the GLF and LEV dialects unverified. Table 1 provides an example of unverified MSA translations of tweets written in the GLF and LEV dialects. The colour-coding highlights translation errors: words in red are those that were not present in the original tweet, while words in orange indicate translation mistakes, such as spelling errors or the use of DA vocabulary. These MSA translations will be verified in this study as explained in Section 4.1.



MSA (Unverified)	GLF Tweet
كيفارجع احب تويتر مثل الاول وخيم فيه 24 ساعه	سعاااااال شصار علئ ألتفاقية الأمنية ؟؟؟ يقولك : عينها زرقت
سعال ماذا حدث في الاتفاق يقولون عينها ازرقت	
سعال شصار علئ ألتفاقية ألأمنية يقولك عينها زرقت	

MSA (Unverified)	LEV Tweet
التغريدة: احنا كان عنا هيك قبل وفاة تيتا	احنا كان عنا هيك قبل وفاة تيتا 💔😔
إذا كان الوضوح من أجل الإفهام، والقوة من أجل التأثير	
نحن كان عندنا هكذا قبل وفاة الجدة	

Table 1: A tweet from GLF and LEV dialects and their unverified MSA translations from Dial2MSA dataset

3.2 Additional Resources

While exploring potentially useful publicly available datasets, we found the following datasets to enrich the training dataset, namely the PADIC (Meftouh et al., 2018), the Multi Arabic Dialect Applications and Resources (MADAR) (Bouamor et al., 2018), the Semantic Textual Similarity (STS) (Al Sulaiman et al., 2022), and the EmiNADI dataset (Khered et al., 2023).

The PADIC (Meftouh et al., 2018) is a multilingual parallel dataset that encompasses sentences from six cities across the LEV and MGR regions, along with corresponding MSA translations. It was developed to improve statistical machine translation between these dialects and MSA.

The MADAR (Bouamor et al., 2018) dataset introduced a multilingual parallel dataset of 25 Arabic city-specific dialects and MSA.

The STS (Al Sulaiman et al., 2022) dataset assesses the semantic similarity between two sentences. It includes translations between EGY and Saudi dialects and MSA.

The EmiNADI (Khered et al., 2023) dataset was created to fill the gap of parallel corpora for the Emirati dialect in NADI 2023 shared task (Abdul-Mageed et al., 2023). It includes MSA translations of Emirati tweets from the training datasets used for NADI 2023 Subtask 1. These translations were produced using the large language model GPT 3.5 Turbo, totalling 2712 translations. Among these, 1000 translations were manually checked by native Arabic speakers to ensure quality.

4 Methodology

4.1 Dial2MSA Verification

This section demonstrates the verification phase, which includes several steps as presented in Figure 2. This process led to the creation of the Dial2MSA-Verified dataset.

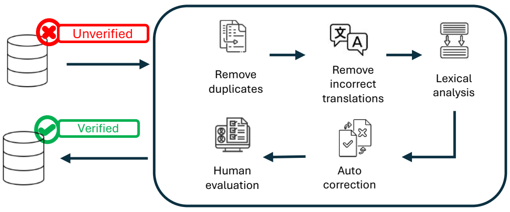
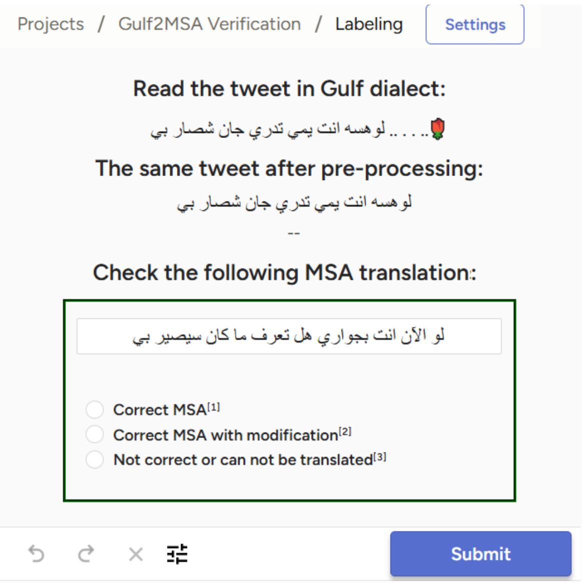


Figure 2: Dial2MSA Verification

Both the unverified GLF and LEV datasets were fed to the cleaning process. Firstly, we removed duplicated samples when a sample had the exact MSA translation. The second step is to remove the incorrectly translated samples. This was conducted by removing samples that included non-Arabic words as well as samples that included dialectal words in the MSA translations. Such words are listed in the research (Mubarak, 2018). The last step is the lexical analysis. This involves the removal of samples that have large different numbers of segments between DA and MSA pairs.

Before the cleaned samples were presented to the annotators, we utilised an Arabic auto-correction tool[2] to correct some of the mistakes automatically. Additionally, we employed GPT-4 via its API[3] to further enhance the correction process. Once these automated steps were completed, the samples were given to human annotators for final verification. We provided them to six native Arabic speakers, three of whom were native speakers of the GLF dialect and three of the LEV dialect. This review process was conducted using Label Studio [4], an open-source online tool that facilitates the annotation and labelling of data. Figure 3 illustrates the human annotation interface in Label Studio.



Read the tweet in Gulf dialect:

🌹.. . . . ..لوهسه انت يمي تدري جان شصار بي

The same tweet after pre-processing:

لوهسه انت يمي تدري جان شصار بي

--

Check the following MSA translation:

لو الآن انت بجواري هل تعرف ما كان سيصير بي

○ Correct MSA[1]
○ Correct MSA with modification[2]
○ Not correct or can not be translated[3]

Figure 3: Human annotation interface in Label Studio

Each annotator had three options: ’correct MSA’, ’correct MSA with modification’, or ’not correct or cannot be translated’. The third option is when the provided MSA translation is in a dialect or if it is too difficult to comprehend or translate. For details

[2] https://pypi.org/project/ar-corrector/

[3] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4/

[4] https://labelstud.io/

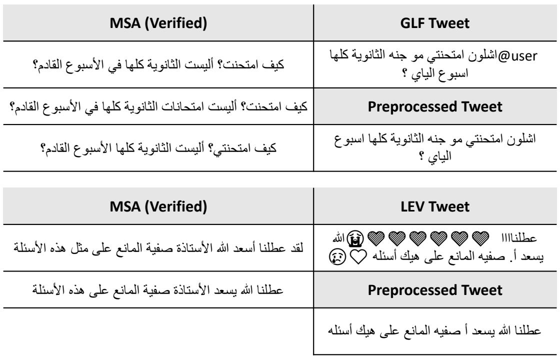
on the annotation guidelines, refer to Appendix A. Table 2 presents the statistics from Dial2MSA-Verified, which includes the original Dial2MSA statistics as well as the updated statistics for GLF and LEV after completing the verification task.

Dialect	Original Tweets	MSA (Task1)	Verified MSA (Task2)	Rem. Tweets	Avg. MSA/ Tweet
EGY	6,000	30,000	16,355	5,565	2.94
MGR	6,000	18,000	7,912	4,953	1.6
LEV	6,000	18,000	8,301	5,319	1.56
GLF	6,000	18,000	12,775	5,354	2.39

Table 2: Statistics for Dial2MSA-Verified corpus after verifying the remaining (Rem.) dialects, specifically GLF and LEV in Dial2MSA (Mubarak, 2018) dataset

4.2 Data Preprocessing and Preparation

Arabic text on social media is usually informal (not standard) and commonly has spelling mistakes, extra characters, diacritical marks, elongations and shortened words. To reduce the noise of such text before applying the Seq2Seq normalisation models, we performed different cleaning and preprocessing methods, such as removing non-Arabic characters, mentions, links, and emojis and dealing with hashtags by including them if only they were written in Arabic. All diacritics were removed, and elongations, in which words contain repeated characters, were stripped. Finally, we removed duplicated samples found after preprocessing before training our models. Table 3 shows an example of tweets in GLF and LEV dialects before and after being preprocessed. It also presents the MSA-verified translations of the tweets.



MSA (Verified)	GLF Tweet
كيف امتحنت؟ أليست الثانوية كلها في الأسبوع القادم؟	@userاشلون امتحنتي مو جنه الثانوية كلها اسبوع الياي ؟
كيف امتحنت؟ أليست امتحانات الثانوية كلها في الأسبوع القادم؟	**Preprocessed Tweet**
كيف امتحنتي؟ أليست الثانوية كلها الأسبوع القادم؟	اشلون امتحنتي مو جنه الثانوية كلها اسبوع الياي ؟

MSA (Verified)	LEV Tweet
لقد عطلنا أسعد الله الأستاذة صفية المانع على مثل هذه الأسئلة	عطلنااا 🖤🖤🖤🖤🖤🖤 😭 الله يسعد أ. صفيه المانع على هيك أسئله 🤍 😢
عطلنا الله يسعد الأستاذة صفية المانع على هذه الأسئلة	**Preprocessed Tweet**
	عطلنا الله يسعد أ صفيه المانع على هيك أسئله

Table 3: The original tweet from GLF and LEV dialects after applying the preprocessed methods and their verified MSA translations

4.3 Dataset Set Up

We collected multiple DA-MSA datasets focusing on four dialects: EGY, GLF, LEV, and MGR. To prepare the Dial2MSA-Verified dataset for model evaluation, we randomly selected 2,000 tweets for each dialect, with multiple MSA references, to test and evaluate our models. The EGY and GLF tweets have three MSA references each, while the LEV and MGR tweets have two MSA references each. From the remaining tweets in the Dial2MSA-Verified dataset, we randomly picked 200 tweets with a single MSA reference for each dialect to serve as a development set. Finally, the remaining tweets have multiple possible MSA references: EGY with 3,365 tweets and 9,099 MSA references, GLF with 3,154 tweets and 6,575 MSA references, LEV with 3,119 tweets and 4,101 MSA references, and MGR with 2,753 tweets and 3,312 MSA references. These remaining samples were combined with additional resources and will be used for training our models. Table 4 shows the training, development and testing datasets. In the Dial2MSA-Verified-test dataset, "R" indicates the number of available MSA references: 2,000 tweets in EGY and GLF have three MSA references each, and 2,000 tweets in LEV and MGR have two MSA references each.

Dataset	**EGY**	**GLF**	**LEV**	**MGR**
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	25,657	27,445	35,525	58,160
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 set up, where Dial2MSA-V (Verified) is used in the training, validation and testing datasets

4.4 Seq2Seq Models

Text-To-Text Transfer Transformer (T5) (Raffel et al., 2020) is an encoder-decoder Transformer-based model designed to support several NLP tasks, including machine translation. For our work, we specifically utilised the second version of **AraT5**[5] model (Nagoudi et al., 2022; Elmadany et al., 2023), which is a fine-tuned variant of T5 explicitly aimed at handling Arabic tasks. Additionally, we employed **mT5** (Xue et al., 2021), and **mT0** (Muennighoff et al., 2023), which are T5-based

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

models trained on a multitude of languages, including Arabic.

The Bidirectional Autoregressive Transformer (BART) (Lewis, 2019) is another model we utilised, which is developed for text generation tasks such as translation. We incorporated two derived models in our evaluation: **AraBART** (Eddine et al., 2022), and **mBART** (Liu, 2020), version mBART-large-50, which supports multiple languages for translation tasks including Arabic.

Furthermore, we used the **M2M100** model (Fan et al., 2021), version M2M100-418M, a multilingual encoder-decoder model created to facilitate many-to-many translation. It was trained on large datasets spanning 100 languages to enable direct translation between various language pairs.

4.5 Training Configurations

We explored two main training approaches: a joint model that integrates data from all regional dialects and an independent model that specialises in translating specific dialects.

Joint Regional Model (J-R): In this setup, we combined all dialect-to-MSA translation pairs from the relevant regions for the four dialects into a single model. The resulting joint model leverages shared linguistic patterns among the dialects and is designed to translate any dialectal text into MSA, regardless of the specific dialect.

Independent Regional Model (I-R): In this configuration, we developed a separate model for each regional dialect. This approach has four models, each trained exclusively to translate text from one specific dialect into MSA. A dialect identification model is used to determine which translation model should be employed for a given text.

4.6 Dialect Identification

We retrained an ensemble of multiple fine-tuned MARBERT (Abdul-Mageed et al., 2021) models with hyperparameter optimisations (Khered et al., 2022) and evaluated the output on the collected datasets (Table 4). More details about the configuration of the ensemble classification model are in (Khered et al., 2022). The results of the best two combination ensemble-MARBERT models and the confusion matrix of the best performing model are in Appendix B.

4.7 Hyperparameter Optimisation

Two Nvidia V100 GPUs were utilised and adhered to the specified configurations; all models were structured to process input and output sequences with a maximum length of 128 tokens. The learning rate was established at 5e-5, and the batch size was configured to 16. The training process was designed to run for a maximum of 20 epochs with early stopping implemented if no improvement was observed on the validation set for 3 consecutive epochs.

5 Evaluation and Results

In this section, we evaluate our proposed models using the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and chrF++ (Popović, 2017) metrics. During training, we validated the models using the BLEU metric on the development set, selecting the checkpoint that achieved the highest BLEU score. For these optimal checkpoints, we report both BLEU and chrF++ scores on the testing set using the SacreBLEU implementation (Post, 2018). This implementation supports multi-reference evaluation for both metrics, providing a comprehensive assessment of model performance. We present results for two configurations: Joint Regional (J-R) and Independent Regional (I-R).

For the I-R configuration, dialect classification is used to select the appropriate translation model. This classification model was evaluated on the testing set of the Dial2MSA-Verified dataset using accuracy and Macro-Average F1 metrics. Additionally, we compare the results obtained when using a single reference file versus multiple reference files for the Dial2MSA-Verified-test dataset to highlight the impact of reference diversity on evaluation.

Table 5 presents the performance of all models under the I-R configuration, with each model evaluated across the four dialects. An average score (Avg) is also provided for each model to summarise overall performance. It can be seen that the AraT5

Model		EGY	GLF	LEV	MGR	Avg
mT0	BLEU	22.87	44.83	34.81	28.55	32.76
	chrF++	45.35	64.98	57.66	53.06	55.26
mT5	BLEU	23.44	45.35	35.12	29.02	33.23
	chrF++	46.65	66.11	59.06	54.56	56.59
AraT5	BLEU	**27.80**	**47.12**	**38.94**	**32.09**	**36.49**
	chrF++	**50.80**	**67.27**	**61.31**	**56.86**	**59.06**
mBART	BLEU	25.38	45.89	37.71	31.29	35.07
	chrF++	48.28	66.61	60.45	56.52	57.96
AraBART	BLEU	25.77	47.05	38.38	31.48	35.67
	chrF++	48.26	66.65	60.31	56.29	57.88
M2M100	BLEU	25.83	37.28	30.48	28.66	30.56
	chrF++	49.38	61.94	55.74	54.19	55.31

Table 5: Performance of different models using I-R configuration

model outperforms other models across all dialects, achieving the highest average BLEU and chrF++ scores of 36.49 and 59.06, respectively. Furthermore, the mBART and AraBART also perform well, with results comparable to those of AraT5.

Similarly, Table 6 presents the performance of all models using the J-R configuration, evaluated across four dialects, along with an average score (Avg) for each model. The results indicate that the AraT5 model outperforms the other models across all four dialects, achieving an average BLEU score of 41.12 and an average chrF++ score of 62.05. Additionally, AraBART shows strong performance, with results comparable to those of AraT5.

Model		EGY	GLF	LEV	MGR	Avg
mT0	BLEU	27.43	46.02	37.30	30.95	35.42
	chrF++	50.77	66.53	60.26	56.23	58.45
mT5	BLEU	27.80	47.12	38.94	32.09	36.49
	chrF++	50.80	67.27	61.31	56.86	59.06
AraT5	BLEU	**30.94**	**53.96**	**45.37**	**34.24**	**41.12**
	chrF++	**52.94**	**70.86**	**65.40**	**58.99**	**62.05**
mBART	BLEU	29.14	49.86	41.15	32.84	38.25
	chrF++	51.75	68.74	62.85	57.71	60.26
AraBART	BLEU	29.87	51.38	43.07	32.95	39.32
	chrF++	52.26	69.49	64.13	58.12	61.00
M2M100	BLEU	22.58	40.88	33.45	27.78	31.17
	chrF++	45.56	62.01	56.17	53.38	54.28

Table 6: Performance of different models using J-R configuration

The overall comparison between the two configurations shows that the J-R configuration outperforms the I-R configuration. This result is due to two reasons. The J-R configuration may benefit from leveraging shared linguistic patterns among similar dialects during training. Moreover, the I-R configuration depends on a dialect classification model to choose the appropriate translation model for each input. Despite the promising results of the dialect identification model (results in Appendix B), it still does not achieve perfect accuracy.

Furthermore. the use of multiple reference translations significantly enhances the evaluation of the model's performance. A single DA sentence can be translated into MSA in multiple forms due to the rich nature of Arabic morphology and syntax. Each translation can preserve the core meaning while using different vocabulary choices and sentence structures. For example, the Levantine dialectal phrase بدي أروح (I want to go), which can be translated into MSA as أرغب بالذهاب, أريد أن أغادر, or سأذهب, all conveying the same essential meaning. With more reference translations, there is a higher likelihood that the model's output will align with at least one reference, leading to a more accurate assessment.

The highest BLEU and chrF++ metrics scores are achieved when evaluating the proposed models on all available references. As shown in Table 7, the performance of the AraT5 model in the J-R configuration improves when evaluated with multiple references. For both EGY and GLF dialects, which have three MSA reference translations each, the model's performance improves when evaluated on all three references. Similarly, for the LEV and MGR dialects, which have two MSA references each, combining both references still results in better scores than using individual ones.

	Refs.	EGY	GLF	LEV	MGR
BLEU	MSA-1	14.92	33.18	32.71	23.42
	MSA-2	14.88	33.35	32.80	23.44
	MSA-3	14.99	33.92	===	===
	MSA-1-2	24.12	46.38	**45.37**	**34.24**
	MSA-2-3	24.48	47.12	===	===
	MSA-1-2-3	**30.94**	**53.96**	===	===
chrF++	MSA-1	41.72	59.97	57.75	52.23
	MSA-2	41.04	60.19	57.75	52.36
	MSA-3	41.46	60.62	===	===
	MSA-1-2	48.91	67.09	**65.40**	**58.99**
	MSA-2-3	48.66	67.33	===	===
	MSA-1-2-3	**52.94**	**70.86**	===	===

Table 7: Comparison of BLEU and chrF++ scores using single vs. multiple MSA references (Refs.) with the J-R AraT5 model

6 Discussion and Analysis

6.1 Dialectal Challenges and Translation Quality

We provide a comprehensive example table showcasing the original tweets, their gold standard translations, and the corresponding AraT5 translations in Appendix C. Our analysis revealed notable issues in normalising Arabic in social media into MSA, particularly due to OOV tokens, many of which stem from non-Arabic origins, as well as unique expressions tied to specific dialects. In the EGY dialect, for example, several English loanwords have adapted meanings that differ from the literal sense of Arabic. The term أوفر, as seen in أنتى أوفر و هو أوفر و كلكم أوفر مش ذنبى أنى بكره, implies "too much" or "over-the-top" in English. However, the model interpreted it as "more available" (its Arabic literal meaning), resulting in translations like

أنت أوفر و هو أوفر و كلكم أوفر ليس ذنبى انى بكره. Similarly, سكتشاتك (derived from "sketch") was misinterpreted as صمتك (your silence) due to morphological resemblance, while الشوز (from "shoes," but translated to سيارة or "car") posed similar challenges.

The J-R-AraT5 model, however, showed some accuracy in translating certain local dialect words. For instance, انزين was correctly rendered as حسن (good) in MSA, اشلون as كيف (how), and جواك in LEV as بداخلك (inside you). Despite these successes, some local expressions remained difficult. For example, اشدعوه was returned unchanged instead of the correct MSA equivalent, ماذا يحدث (what is happening), and the LEV phrase نيالك (lucky you) could not be accurately translated due to its unique connotation.

Additionally, the model showed an ability to handle other Arabised English terms frequently seen on social media. Words such as الفولورز were appropriately translated to متابعين (followers), and الفايس to فايسبوك (Facebook). It also successfully translated رتويت as اعادة تغريدة (retweet), demonstrating its adaptability to social media language.

MGR dialect posed a different set of challenges, particularly due to lexical and conjugation differences from other dialects. For example, كاتعايرو, meaning "insulting," was often misinterpreted as something related to work (عمل), while واسم (meaning "what is up" in Algerian dialect) was incorrectly translated to اسم (name). The term البيام (referring to the French BEM exam) also created difficulties, as it appears in Arabic script but is inherently non-Arabic.

LEV and EGY dialects featured unique dialectal words that the model struggled to translate accurately, even though direct MSA equivalents exist. Words like شخابيط (doodles), امبارح (yesterday), and انتخة (laziness) were challenging for the model, perhaps due to morphological or contextual ambiguities that made it difficult for the model to identify the correct translations.

6.2 Model Performance and Recommendations

While the model performed effectively with some dialect-specific terms, it often struggled with borrowed words and region-specific vocabulary across all dialects. Improvements in translation quality could be achieved by expanding dialect-specific datasets to include common foreign-origin terms, as well as by integrating context-sensitive embeddings to reduce ambiguity for polysemous words. Additionally, applying more nuanced preprocessing techniques could help account for regional lexical and morphological variations, enabling models to capture the linguistic richness and contextual relevance of Arabic dialects.

7 Conclusion and Future Work

This work introduced Dial2MSA-Verified, an extension of the Dial2MSA dataset that involves the verification of previously unverified dialects. We enriched the training data by incorporating Seq2Seq datasets from various domains. We conducted a comprehensive model evaluation using multi-reference evaluation, demonstrating improved performance compared to single-reference evaluations. Our findings indicate that models trained in the J-R configuration outperformed those in the I-R configuration. This improvement is due to the inherent similarities between dialects, allowing dialects to be learned from one another. Additionally, the I-R configuration relied on dialect identification for model selection, which affected translation performance. Overall, AraT5 outperformed other models, achieving an average BLEU score of 41.12 and a chrF++ score of 62.05.

In future work, we plan to expand the training data, focusing on the social media domain, as the limited availability remains an obstacle. Additionally, we plan to explore the possibility of improving the normalisation performance by leveraging more advanced models, data augmentation techniques and transfer learning techniques.

8 Limitations

While the Dial2MSA-Verified dataset offers comprehensive coverage of multiple dialectal regions, it still lacks representation for other Arabic dialects, such as Sudanese and Yemeni dialects. This gap may limit the model's ability to generalise effectively across all Arabic-speaking regions. Moreover, models trained with Seq2Seq datasets from varied domains might experience difficulties when applied to domain-specific texts, potentially affecting translation accuracy in social media contexts. Lastly, the reliance on dialect identification for model selection in some configurations poses a limitation, as incorrect identification can impact translation performance.

9 Ethical Considerations

This study adhered to ethical guidelines by ensuring data confidentiality and compliance with data protection regulations. Datasets were anonymised, and annotators provided informed consent for voluntary participation. Measures were taken to minimise potential biases by selecting diverse dialectal data and involving annotators from different dialect regions to ensure fairness and accuracy in data verification.

Acknowledgments

We thank the annotators who verified the two dialects in the Dial2MSA-Verified dataset. Additionally, we acknowledge the University of Manchester for providing access to their High-Performance Computing (HPC) cluster, the Computational Shared Facility (CSF), which was used for the training process. We extend our gratitude to the Ministry of Education in Saudi Arabia and King Abdulaziz University for their continued support.

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, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. Nadi 2023: The fourth nuanced arabic dialect identification shared task. *Preprint*, arXiv:2310.16117.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. Nadi 2024: The fifth nuanced arabic dialect identification shared task. *Preprint*, arXiv:2407.04910.

G. H. Al-Gaphari and M. Al-Yadoumi. 2010. A method to convert sana'ani accent to modern standard arabic. *International Journal of Information Science and Management (IJISM)*, 8(1):39–49.

Roqayah Al-Ibrahim and Rehab M Duwairi. 2020. Neural machine translation from jordanian dialect to modern standard arabic. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 173–178. IEEE.

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.

Abubakar Aliero, Sulaimon Bashir, Hamzat Aliyu, Amina Tafida, Bashar Kangiwa, and Nasiru Dankolo. 2023. Systematic review on text normalization techniques and its approach to non-standard words. *International Journal of Computer Applications*, 185:975–8887.

Manan AlMusallam and Samar Ahmad. 2024. Alson at NADI 2024 shared task: Alson - a fine-tuned model for Arabic dialect translation. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 764–768, Bangkok, Thailand. Association for Computational Linguistics.

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

Maryam Aminian, Mahmoud Ghoneim, and Mona Diab. 2014. Handling oov words in dialectal arabic to english machine translation. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 99–108.

Laith H. Baniata, Isaac. K. E. Ampomah, and Seyoung Park. 2021. A transformer-based neural machine translation model for arabic dialects that utilizes subword units. *Sensors*, 21(19).

Nouhaila Bensalah, Habib Ayad, Abdellah Adib, and Abdelhamid Ibn El Farouk. 2021. Transformer model and convolutional neural networks (cnns) for arabic to english machine translation. In *International Conference On Big Data and Internet of Things*, pages 399–410. Springer.

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*.

Mohammed ElAmine Chennafi, Hanane Bedlaoui, Abdelghani Dahou, and Mohammed A. A. Al-qaness. 2022. Arabic aspect-based sentiment classification using seq2seq dialect normalization and transformers. *Knowledge*, 2(3):388–401.

Rehab M. Duwairi. 2015. Sentiment analysis for dialectical arabic. In *2015 6th International Conference on Information and Communication Systems (ICICS)*, pages 166–170.

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. *arXiv preprint arXiv:2203.10945*.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Octopus: A multitask model and toolkit for arabic natural language generation. *Preprint*, arXiv:2310.16127.

Ashraf Hatim Elneima, AhmedElmogtaba Abdelmoniem Ali Abdelaziz, and Kareem Darwish. 2024. Osact6 dialect to msa translation shared task overview. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 93–97.

GÜLŞEN ERYİĞİT and DİLARA TORUNOĞLU-SELAMET. 2017. Social media text normalization for turkish. *Natural Language Engineering*, 23(6):835–875.

Mohamed Atta Faheem, Khaled Tawfik Wassif, Hanaa Bayomi, and Sherif Mahdy Abdou. 2024. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of egyptian dialect to modern standard arabic translation. *Scientific Reports*, 14(1):2265.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Mahmoud Ghoneim and Mona Diab. 2013. Multiword expressions in the context of statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1181–1187.

Salwa Hamada and Reham Marzouk. 2018. *Developing a Transfer-Based System for Arabic Dialects Translation*, pages 121–138.

Ahmed Hamdi, Rahma Boujelbane, Nizar Habash, and Alexis Nasr. 2013. The effects of factorizing root and pattern mapping in bidirectional tunisian-standard arabic machine translation. In *Proceedings of Machine Translation Summit XIV: Papers*.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing Management*, 56(2):262–273. Advance Arabic Natural Language Processing (ANLP) and its Applications.

Mohamed Osman Hegazi, Yasser Al-Dossari, Abdullah Al-Yahy, Abdulaziz Al-Sumari, and Anwer Hilal. 2021. Preprocessing arabic text on social media. *Heliyon*, 7(2).

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 Salem 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.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Y Liu. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

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

Claudia Matos Veliz, Orphée De Clercq, and Veronique Hoste. 2021. Is neural always better? smt versus nmt for dutch text normalization. *Expert Systems with Applications*, 170:114500.

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

Hamdy Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. *OSACT*, 3:49.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, and Teven et al. Le Scao. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

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.

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, Belgium, Brussels. Association for Computational Linguistics.

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.

Sandra Rizkallah, Amir Atiya, Hossam ElDin Mahgoub, and Momen Heragy. 2018. Dialect versus msa sentiment analysis. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, pages 605–613. Springer.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of arabic dialects in social media. SoMeRA '14, page 35–40, New York, NY, USA. Association for Computing Machinery.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal arabic to english. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6.

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.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties*, pages 10–21.

Wael Salloum and Nizar Habash. 2012. Elissa: A dialectal to standard arabic machine translation system. In *International Conference on Computational Linguistics*.

Wael Salloum and Nizar Habash. 2013. Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358.

Wael Salloum and Nizar Habash. 2014. Adam: Analyzer for dialectal arabic morphology. *Journal of King Saud University-Computer and Information Sciences*, 26(4):372–378.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Pamela Shapiro and Kevin Duh. 2019. Comparing pipelined and integrated approaches to dialectal Arabic neural machine translation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 214–222, Ann Arbor, Michigan. Association for Computational Linguistics.

Amel Slim and Ahlem Melouah. 2024. Low resource arabic dialects transformer neural machine translation improvement through incremental transfer of shared linguistic features. *Arabian Journal for Science and Engineering*, pages 1–17.

Amel Slim, Ahlem Melouah, Usef Faghihi, and Khouloud Sahib. 2022. Improving neural machine translation for low resource algerian dialect by transductive transfer learning strategy. *Arabian Journal for Science and Engineering*, 47(8):10411–10418.

Ridouane Tachicart and Karim Bouzoubaa. 2014. A hybrid approach to translate moroccan arabic dialect. In *2014 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14)*, pages 1–5.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jezia Zakraoui, Moutaz Saleh, Somaya Al-Maadeed, and Jihad Mohamed Alja'am. 2021. Arabic machine translation: A survey with challenges and future directions. *IEEE Access*, 9:161445–161468.

Randa Zarnoufi, Hamid Jaafar, and Mounia Abik. 2020. Machine normalization: Bringing social media text from non-standard to standard form. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(4).

A Annotation Guidelines

Overview

Thank you for agreeing to assist us with verifying Modern Standard Arabic (MSA) translations. One at a time, you will be presented with tweets in the [Gulf/Levantine] dialect and their MSA translations. Your task is to verify these MSA translations.

Instructions

You will be using a tool called Label Studio. Before starting, you will be shown examples of correct translations and tested to ensure you are prepared for this task. This preparation helps to maintain high quality in the work.

Your Role

You will verify the MSA translations of [Gulf/Levantine] dialect tweets. For each translation, you have the following three choices:

- **Correct MSA**: Select this option if the translation is accurate and requires no changes.
- **Correct MSA with Modification**: Choose this option if the translation is partially correct and requires corrections. Please specify the corrected translation in the textbox.
- **Not Correct or Cannot be Translated**: Use this option if the translation contains significant errors, remains in dialect, or is too difficult to understand or translate.

When reviewing translations, ensure that the MSA translation accurately conveys the original meaning of the tweet. Check for spelling and grammatical correctness as well as proper sentence structure in the MSA translation.

B DA Identification Results and Model Performance Metrics

Model	Accuracy	Macro-Average F1
Ensemble-MARBERT-1	96.95	96.942
Ensemble-MARBERT-2	96.224	96.215

Table 8: Results for DA identification based on the ensemble of best two combinations of MARBERT models

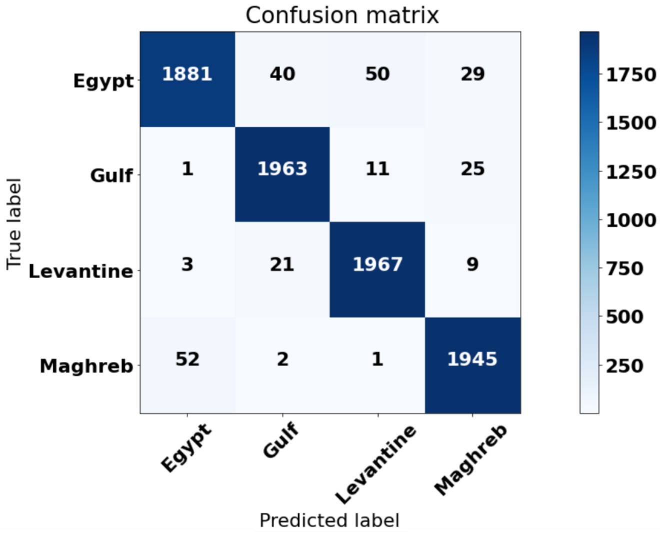
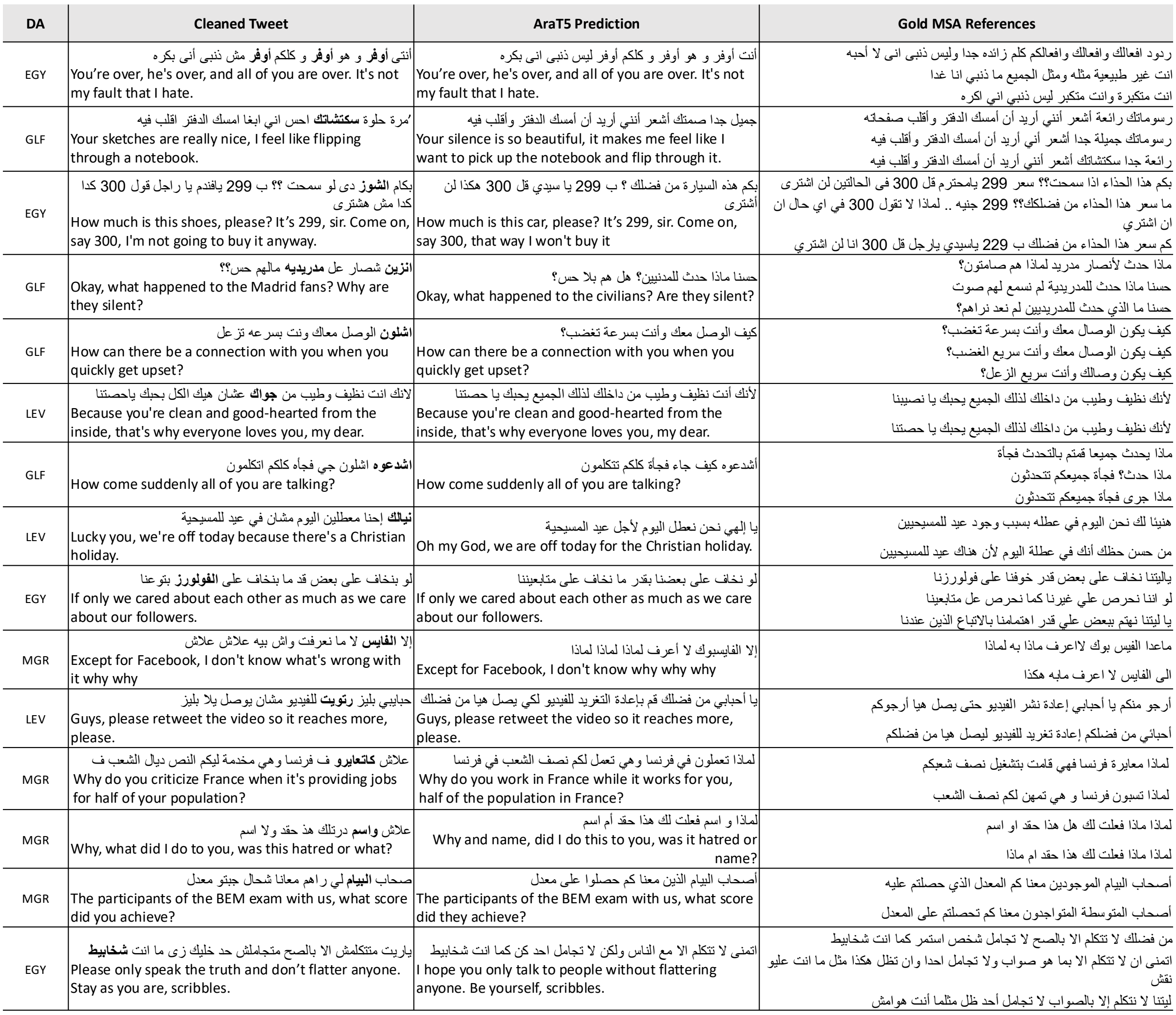


Figure 4: Confusion matrix for the best-performing ensemble model of DA identification

C Comparative Analysis of J-R AraT5 Predictions and Gold MSA References



DA	Cleaned Tweet	AraT5 Prediction	Gold MSA References
EGY	أنتى **أوفر** و هو **أوفر** و كلكم **أوفر** مش ذنبى أنى بكره You’re over, he's over, and all of you are over. It's not my fault that I hate.	أنت أوفر و هو أوفر و كلكم أوفر ليس ذنبى انى بكره You’re over, he's over, and all of you are over. It's not my fault that I hate.	ردود افعالك وافعالكم كلم زائده جدا وليس ذنبى انى لا أحبه انت غير طبيعية مثله ومثل الجميع ما ذنبي انا غدا انت متكبرة وانت متكبر ليس ذنبي اني اكره
GLF	’مرة حلوة **سكتشاتك** احس اني ابغا امسك الدفتر اقلب فيه Your sketches are really nice, I feel like flipping through a notebook.	جميل جدا صمتك أشعر أنني أريد أن أمسك الدفتر وأقلب فيه Your silence is so beautiful, it makes me feel like I want to pick up the notebook and flip through it.	رسوماتك رائعة أشعر أنني أريد أن أمسك الدفتر وأقلب صفحاته رسوماتك جميلة جدا أشعر أني أريد أن أمسك الدفتر وأقلب فيه رائعة جدا سكتشاتك أشعر أنني أريد أن أمسك الدفتر وأقلب فيه
EGY	بكام **الشوز** دى لو سمحت ؟؟ ب 299 يافندم يا راجل قول 300 كدا كدا مش هشترى How much is this shoes, please? It’s 299, sir. Come on, say 300, I'm not going to buy it anyway.	بكم هذه السيارة من فضلك ؟ ب 299 يا سيدي قل 300 هكذا لن أشترى How much is this car, please? It’s 299, sir. Come on, say 300, that way I won't buy it	بكم هذا الحذاء اذا سمحت؟؟ سعر 299 يامحترم قل 300 فى الحالتين لن اشترى ما سعر هذا الحذاء من فضلك؟؟ 299 جنيه .. لماذا لا تقول 300 في اي حال ان اشتري كم سعر هذا الحذاء من فضلك ب 229 ياسيدي يارجل قل 300 انا لن اشتري
GLF	**انزين** شصار عل **مدريديه** مالهم حس؟؟ Okay, what happened to the Madrid fans? Why are they silent?	حسنا ماذا حدث للمدنيين؟ هل هم بلا حس؟ Okay, what happened to the civilians? Are they silent?	ماذا حدث لأنصار مدريد لماذا هم صامتون؟ حسنا ماذا حدث للمدريدية لم نسمع لهم صوت حسنا ما الذي حدث للمدريديين لم نعد نراهم؟
GLF	**اشلون** الوصل معاك ونت بسرعه تزعل How can there be a connection with you when you quickly get upset?	كيف الوصل معك وأنت بسرعة تغضب؟ How can there be a connection with you when you quickly get upset?	كيف يكون الوصال معك وأنت بسرعة تغضب؟ كيف يكون الوصال معك وأنت سريع الغضب؟ كيف يكون وصالك وأنت سريع الزعل؟
LEV	لانك انت نظيف وطيب من **جواك** عشان هيك الكل بحبك ياحصتنا Because you're clean and good-hearted from the inside, that's why everyone loves you, my dear.	لأنك أنت نظيف وطيب من داخلك لذلك الجميع يحبك يا حصتنا Because you're clean and good-hearted from the inside, that's why everyone loves you, my dear.	لأنك نظيف وطيب من داخلك لذلك الجميع يحبك يا نصيبنا لأنك نظيف وطيب من داخلك لذلك الجميع يحبك يا حصتنا
GLF	**اشدعوه** اشلون جي فجأه كلكم اتكلمون How come suddenly all of you are talking?	أشدعوه كيف جاء فجأة كلكم تتكلمون How come suddenly all of you are talking?	ماذا يحدث جميعا قمتم بالتحدث فجأة ماذا حدث؟ فجأة جميعكم تتحدثون ماذا جرى فجأة جميعكم تتحدثون
LEV	**نيالك** إحنا معطلين اليوم مشان في عيد للمسيحية Lucky you, we're off today because there's a Christian holiday.	يا إلهي نحن نعطل اليوم لأجل عيد المسيحية Oh my God, we are off today for the Christian holiday.	هنيئا لك نحن اليوم في عطله بسبب وجود عيد للمسيحيين من حسن حظك أنك في عطلة اليوم لأن هناك عيد للمسيحيين
EGY	لو بنخاف على بعض قد ما بنخاف على **الفولورز** بتوعنا If only we cared about each other as much as we care about our followers.	لو نخاف على بعضنا بقدر ما نخاف على متابعيننا If only we cared about each other as much as we care about our followers.	ياليتنا نخاف على بعض قدر خوفنا على فولورزنا لو اننا نحرص علي غيرنا كما نحرص عل متابعينا يا ليتنا نهتم ببعض علي قدر اهتمامنا بالاتباع الذين عندنا
MGR	إلا **الفايس** لا ما نعرفت واش بيه علاش علاش Except for Facebook, I don't know what's wrong with it why why	إلا الفايسبوك لا أعرف لماذا لماذا لماذا Except for Facebook, I don't know why why why	ماعدا الفيس بوك لااعرف ماذا به لماذا الى الفايس لا اعرف مابه هكذا
LEV	حبايبي بليز **رتويت** للفيديو مشان يوصل يلا بليز Guys, please retweet the video so it reaches more, please.	يا أحبابي من فضلك قم بإعادة التغريد للفيديو لكي يصل هيا من فضلك Guys, please retweet the video so it reaches more, please.	أرجو منكم يا أحبابي إعادة نشر الفيديو حتى يصل هيا أرجوكم أحبائي من فضلكم إعادة تغريد للفيديو ليصل هيا من فضلكم
MGR	علاش **كاتعايرو** ف فرنسا وهي مخدمة ليكم النص ديال الشعب ف Why do you criticize France when it's providing jobs for half of your population?	لماذا تعملون في فرنسا وهي تعمل لكم نصف الشعب في فرنسا Why do you work in France while it works for you, half of the population in France?	لماذا معايرة فرنسا فهي قامت بتشغيل نصف شعبكم لماذا تسبون فرنسا و هي تمهن لكم نصف الشعب
MGR	علاش **واسم** درتلك هذ حقد ولا اسم Why, what did I do to you, was this hatred or what?	لماذا و اسم فعلت لك هذا حقد أم اسم Why and name, did I do this to you, was it hatred or name?	لماذا ماذا فعلت لك هل هذا حقد او اسم لماذا ماذا فعلت لك هذا حقد ام ماذا
MGR	صحاب **البيام** لي راهم معانا شحال جبتو معدل The participants of the BEM exam with us, what score did you achieve?	أصحاب البيام الذين معنا كم حصلوا على معدل The participants of the BEM exam with us, what score did they achieve?	أصحاب البيام الموجودين معنا كم المعدل الذي حصلتم عليه أصحاب المتوسطة المتواجدون معنا كم تحصلتم على المعدل
EGY	ياريت متتكلمش الا بالصح متجاملش حد خليك زى ما انت **شخابيط** Please only speak the truth and don’t flatter anyone. Stay as you are, scribbles.	اتمنى لا تتكلم الا مع الناس ولكن لا تجامل احد كن كما انت شخابيط I hope you only talk to people without flattering anyone. Be yourself, scribbles.	من فضلك لا تتكلم الا بالصح لا تجامل شخص استمر كما انت شخابيط اتمنى ان لا تتكلم الا بما هو صواب ولا تجامل احدا وان تظل هكذا مثل ما انت عليو نقش ليتنا لا نتكلم إلا بالصواب لا تجامل أحد ظل مثلما أنت هوامش

Table 9: Comparison of J-R AraT5 predictions and gold MSA references across the four dialects in the Dial2MSA-Verified dataset. Bold words highlight challenging dialect terms discussed in the paper’s discussion section.