DialectNLU at NADI 2023 Shared Task: Transformer Based Multitask Approach Jointly Integrating Dialect and Machine Translation Tasks in Arabic

Hariram Veeramani
Department of Electrical
and Computer Engineering,
UCLA, USA
hariram@ucla.edu

Surendrabikram Thapa
Department of Computer
Science, Virginia Tech,
Blacksburg, USA
sbt@vt.edu

Usman Naseem
College of Science and
Engineering, James Cook
University, Australia
usman.naseem@jcu.edu.au

Abstract

With approximately 400 million speakers worldwide, Arabic ranks as the fifth most-spoken language globally, necessitating advancements in natural language processing. This paper describes the approaches employed for the subtasks outlined in the Nuanced Arabic Dialect Identification (NADI) task at EMNLP 2023. We employ an ensemble of two Arabic language models for the first subtask involving closed country-level dialect identification classification. Similarly, for the second subtask, focused on closed dialect to Modern Standard Arabic (MSA) machine translation, our approach combines sequence-to-sequence models trained on an Arabic-specific dataset. Our team ranks 10th and 3rd on subtask 1 and subtask 2, respectively.

1 Introduction

The Arabic language, with approximately 400 million speakers across the globe, stands as the fifth most widely spoken language worldwide (Mohammed Ameen et al., 2023). Its vast linguistic diversity, rooted in rich historical and regional variations, necessitates continuous advancements in the field of natural language processing (NLP) (Abdul-Mageed et al., 2020). Within the scope of this linguistic diversity, Modern Standard Arabic (MSA) serves as the standardized form of the language, fostering communication across Arabic-speaking regions. However, coexisting with MSA are numerous dialects, each bearing its unique linguistic features and nuances (Abdul-Mageed et al., 2021b).

Arabic encompasses a wide array of languages and language variations, with some of them lacking mutual intelligibility (Abdul-Mageed et al., 2022a) (Veeramani et al., 2023d,c,b). Despite this diversity, Arabic is frequently misconceived as a single, uniform language. Thus, identifying these different dialects plays a pivotal role in the realm of Arabic language understanding, primarily due to the contextual intricacies they introduce (Salameh et al., 2018). Dialect identification serves as the bedrock for a multitude of NLP applications, enabling accurate language understanding, effective communication, sentiment analysis, and sociolinguistic insights (Malmasi et al., 2015; Veeramani et al., 2023e,a,f). Furthermore, dialect classification preserves and celebrates the rich linguistic diversity encapsulated within the Arabic language landscape (Zaidan and Callison-Burch, 2014; Salameh et al., 2018).

Similarly, machine translation, in particular, holds profound significance within this rich Arabic linguistic landscape (Kchaou et al., 2023). Bridging the gap between dialects and the standardized form of the language, MSA, and machine translation facilitates seamless communication across Arabic-speaking communities (Al-Ibrahim and Duwairi, 2020). In an interconnected world where communication knows no borders, machine translation becomes the vital bridge that transcends linguistic differences (Ameur et al., 2020).

In this paper, we address the pressing need for advancements in Arabic dialect identification and machine translation. Specifically, we present our contributions to the Nuanced Arabic Dialect Identification (NADI) task (Abdul-Mageed et al., 2023) at 1st ArabicNLP colacted with EMNLP 2023. Our work centers on two crucial subtasks:

- **Closed Country-Level Dialect Identification**: To tackle this subtask, we leverage an ensemble of two Arabic language models, harnessing the power of natural language processing to classify dialects at the country level.
- **Closed Dialect to MSA Machine Translation**: For this subtask, we employ a combination of sequence-to-sequence models, all meticulously trained on an Arabic-specific dataset. Our goal is to enhance the translation accuracy of Arabic dialects into the standardized MSA, thereby promoting effective

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 614–619
December 7, 2023 ©2023 Association for Computational Linguistics

cross-dialect communication.

This paper explains our systems in detail, offering comprehensive insights into our approach, the rationale behind our methodology, a thorough analysis of our results, and valuable insights derived from our findings.

2 Task Descriptions

We submitted results for the first two out of three subtasks.

Subtask 1: This subtask involves identifying the dialect of a given text, with a particular emphasis on Arabic dialects that lack well-defined linguistic conventions and structures. The evaluation metric for this subtask is the macro-averaged F1-score.

Subtask 2: This subtask entails translating non-MSA dialects into Modern Standard Arabic (MSA) across four specified dialects. Evaluation is based on the BLEU score.

3 Dataset

Subtask 1 focuses on informal Twitter discourse featuring languages from Arabic-speaking nations, including Qatar, Syria, Libya, Yemen, Kuwait, Morocco, UAE, Jordan, Palestine, Tunisia, Saudi Arabia, Egypt, Iraq, Algeria, Bahrain, Sudan, Lebanon, and Oman. The dataset, NADI-2023-TWT, consists of 18,000 training tweets (1,000 per country), 1,800 tweets in the dev dataset (100 per country), and 3,600 tweets in the test dataset. I

Subtask 2, on the other hand, provides a manually curated dataset focusing on four urban dialects: Egyptian, Emirati, Jordanian, and Palestinian. The training data primarily originates from the MADAR-parallel corpus (Bouamor et al., 2018), comprising 40,000 sentences. The test set consists of 2,000 sentences (500 from each dialect), while the dev set comprises 400 sentences (100 from each of the four dialects). It is important to note that we did not use any external data or augmentation.

4 System Description

4.1 Subtask 1

In addressing the classification problem of subtask 1, we employ a strategic combination of ARBERT and MARBERT (Abdul-Mageed et al., 2021a). ARBERT is trained on Modern Standard Arabic (MSA) data, whereas MARBERT specializes in learning from the informal dialects commonly found in Twitter data. This selection is grounded in the recognition that Arabic encompasses diverse language styles. ARBERT excels in comprehending formal Arabic, particularly MSA rules, making it proficient in handling general aspects of the language. On the other hand, MARBERT, fine-tuned on Twitter's informal dialects, adeptly captures the nuances of day-to-day expressions. We perform the combination with various strategies.

4.1.1 Max-voting Ensemble

As a first strategy, at the logit level, we implemented a weighted ensemble approach. ARBERT was assigned a weight of 0.4, while MARBERT received a weight of 0.6. This weighting strategy was adopted to optimize the ensemble's performance by capitalizing on the unique strengths of each model. The higher weight assigned to MARBERT reflects its proficiency in capturing informal nuances from Twitter data, ensuring robust and accurate dialect classification across diverse Arabic language variations encountered in online contexts.

4.1.2 Fusion Representation Technique

In our second strategy, we fuse the hidden representations of both ARBERT and MARBERT models to leverage their complementary strengths (Abdul-Mageed et al., 2022b), enhancing the model's ability to capture nuanced dialect features. Throughout this paper, this representation of dimensions 2×768 will be referred to as fusion representation. Following this, we incorporate a dropout layer (DO) to enhance the model's performance. This approach has proven to be the most effective model for subtask 1. Additionally, we also experiment by incorporating a label-aware technique (LAT) by appending the respective label to the beginning of the text input.

4.2 Subtask 2

For the machine translation challenge of subtask 2, we applied various models, including AraBART, AraT5 (base), and AraT5 (base-1024). We also choose an ensemble approach that combines dialect classifier and AraBART in two settings. This ensemble strategy was selected because each model has unique strengths and weaknesses. By merging them, we effectively mitigate these weaknesses, resulting in more precise and robust translations. The standalone models and ensemble approach are

Model	Dev Dataset				Test Dataset (F1 score)
	F1-score	Precision	Recall	Accuracy	
ARBERT	75.5	76.1	75.34	75.38	73.16
MARBERT	76.3	76.9	76.0	76.0	73.25
Max-voting Ensemble	77.5	79.6	77.2	77.2	77.19
Fusion Representation	79.75	80.97	79.66	79.66	79.06
Fusion Representation + DO + LAT	79.83	79.81	79.75	79.81	79.06
Fusion Representation + DO	**80.55**	**80.98**	**80.44**	**80.44**	**80.56**

Table 1: Performance of various model combinations for task 1 on dev and test dataset. All scores reported are macro-averaged scores. The abbreviations are introduced in section 4.1.2.

defined below.

4.2.1 Standalone Models

We use three different standalone models for our machine translation task. The models and the motivation for using them are explained below:

AraT5 (base): This sequence-to-sequence model is pre-trained on a substantial Arabic text corpus, encompassing Modern Standard Arabic (MSA) and Arabic dialects. This extensive training gives AraT5 (base) a profound understanding of Arabic grammar and vocabulary, prerequisites for accurate translation. AraT5 (base)[1] is also trained using a denoising-based pre-training methodology, which enhances its capacity to handle noisy data—an invaluable trait for machine translation, where source texts may contain errors or typos.

AraBART: This is another powerful machine translation model, albeit trained on a comparatively smaller corpus of Arabic text (Kamal Eddine et al., 2022). Because of its modeling architecture, AraT5 (base)[2] may have a superior grasp of Arabic grammar and vocabulary. Additionally, AraBART undergoes training with a distinct denoising-based pre-training method, potentially better suited for processing noisy data compared to AraT5.

AraT5 (base-1024): This variant of AraT5 benefits from training on an even larger corpus of text and boasts a more extensive vocabulary compared to AraT5 (base). Its broader lexicon and nuanced understanding of the Arabic language make AraT5 (base-1024)[3] particularly good at capturing subtleties in translation. Moreover, AraT5 (base-1024) features an extended sequence length and faster convergence during fine-tuning, expediting the training process.

4.2.2 Ensemble Approach

In our ensemble approach for Subtask 2, we leverage AraBART, a sequence-to-sequence classifier, in two distinct settings. In both settings, we initially fine-tuned AraBART for the dialect classification task. Subsequently, we remove the classifier head and perform an additional fine-tuning phase, focusing on sequence-to-sequence translation. The key difference between the two settings lies in the learning scheduler employed for AraBART. One setting utilizes a 'linear' learning scheduler, while the other adopts a 'cosine' learning scheduler. This variation in learning scheduler choice allows us to explore different training dynamics. When determining which translation to use in the ensemble, we opt for the model that excels in the specific task (Kanagasabai et al., 2023), dialect classification. This approach helps optimize the overall performance of the ensemble in accurately translating non-MSA dialects into Modern Standard Arabic.

5 Results

Table 1 presents a comprehensive evaluation of various models and model combinations used for Subtask 1, focusing on dialect classification. It includes key metrics such as F1-score, precision, recall, and accuracy for both the development (Dev) and Test datasets. The Dev dataset serves as a validation set for fine-tuning, while the Test dataset represents the models' expected performance in real-world scenarios. Notably, ARBERT and MARBERT exhibit strong dialect classification capabilities on the Dev dataset, achieving scores of 75.5 and 76.3, respectively. The Max-voting Ensemble strategy enhances performance, yielding an F1-score of 77.5. Fusion Representation further elevates dialect identification with a score of 79.75. The Fusion Representation model with Dropout and Label-aware Training (LAT) attains an even higher performance on the Dev dataset, registering an F1-score of 79.83.

[1] https://huggingface.co/UBC-NLP/AraT5-base

[2] https://huggingface.co/moussaKam/AraBART

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

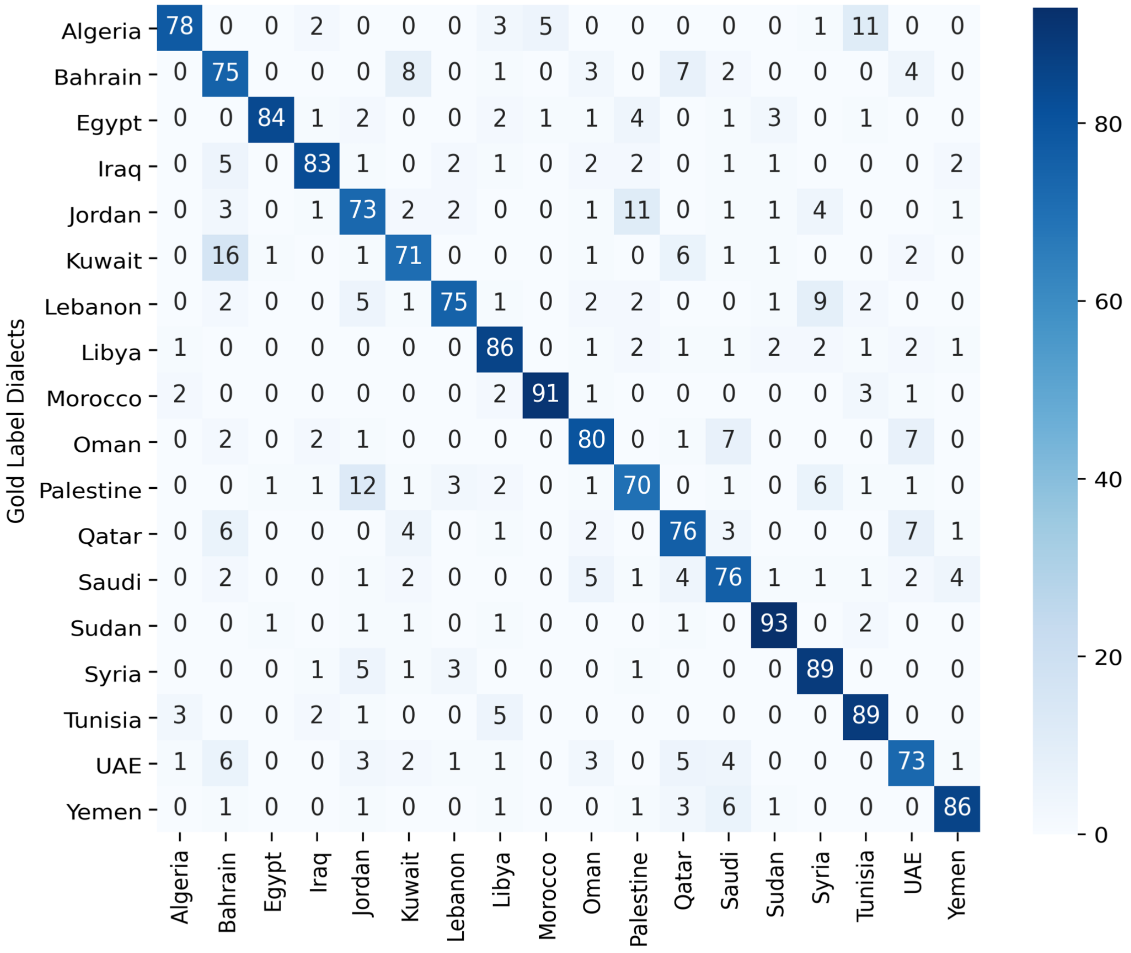

Figure 1: Confusion matrix for our best-performing model for country-wise dialect classification (task 1). We use the dev set data for this analysis.

The Fusion Representation model with Dropout stands out as the top-performing model, achieving an impressive F1-score of 80.55. A similar trend can be seen with the test dataset as well. Our fusion representation model performs the best in the test dataset with an F1-score of 80.56.

To provide deeper insights into our models' performance, Figure 1 presents confusion matrices for all 18 dialects in Subtask 1. These matrices offer a detailed breakdown of classification results, shedding light on how well our models identify each dialect. Notably, Sudanese and Moroccan dialects exhibit a strong classification, while some challenges persist in accurately classifying Palestinian and Kuwaiti dialects. The confusion matrix serves as a valuable tool for understanding model performance in specific dialects and identifying areas for improvement, further enriching our analysis of dialect classification results.

Similarly, Table 2 comprehensively compare model performance in Subtask 2, focusing on the machine translation of non-Modern Standard Arabic (MSA) dialects into MSA. The models assessed include AraT5 (base), which achieves a BLEU score of 0.54 on the evaluation dataset and 0.014 on the test dataset, indicating translation challenges. AraT5 (base-1024) exhibits improvement with BLEU scores of 1.03 and 0.07 on the evaluation and test datasets, respectively. Among standalone models, AraBART excels with high BLEU

Model	Eval BLEU	Test BLEU
AraT5 (base)	0.54	0.014
AraT5 (base-1024)	1.03	0.07
AraBART	12.01	13.42
Ensemble Approach	**12.9**	**13.43**

Table 2: Performance of various standalone models along with our ensemble approach for machine translation subtask. Overall BLEU scores are presented for both eval and test datasets.

scores, achieving 12.01 on the evaluation dataset and 13.42 on the test dataset, showcasing its proficiency in accurate dialect translation. Most notably, our novel ensemble approach outperforms individual standalone models, achieving the highest BLEU scores of 12.9 on the evaluation dataset and 13.43 on the test dataset, highlighting the efficacy of ensemble techniques in enhancing machine translation quality for Arabic dialects.

In Subtask 1, focused on dialect identification and Subtask 2, addressing machine translation, ensemble techniques (fusion-level or decision-level) have consistently demonstrated outstanding performance. By strategically combining multiple models, we have harnessed the collective strengths of various standalone models to achieve remarkable results. This underscores the pivotal role of ensemble methodologies in enhancing the accuracy and robustness of Arabic dialect identification and machine translation, reaffirming their effectiveness across diverse linguistic challenges.

6 Conclusion

In conclusion, our participation in the Nuanced Arabic Dialect Identification (NADI) task at EMNLP 2023 has demonstrated the effectiveness of innovative approaches in addressing the intricate challenges posed by Arabic dialect identification and machine translation. With its diverse linguistic landscape, Arabic presents a unique and formidable set of hurdles for natural language processing tasks. The high performance of ensemble strategies that involve carefully combining various models has showcased remarkable achievements in dialect classification and machine translation, underlining the power of ensemble techniques. Furthermore, our contributions extend beyond performance metrics, encompassing comprehensive system descriptions, model rationale, and insights from experimentation. These approaches pave the way for tackling the multi-aspects challenges of Arabic NLP forward.

Ethics Statement

It is important to acknowledge that this research does not include a comprehensive assessment of potential bias in the models deployed. Before real-world applications, models should undergo thorough bias assessments to ensure fairness and equity in their predictions. We encourage future research and practitioners to consider bias assessments as an integral part of deploying these models in practical settings, emphasizing ethical AI practices and responsible AI development.

Limitations

While our approaches have shown promising results, several limitations are worth noting. First, our ensemble strategies, while effective, are computationally intensive and require substantial resources. Implementing these approaches at scale may pose challenges in resource-constrained environments. Second, our models' performance may be influenced by the availability and quality of training data. The scarcity of annotated data for some Arabic dialects could impact the generalization of our models. Additionally, our current strategies primarily focus on closed-track evaluations; extending them to open-domain scenarios remains an avenue for future exploration. Finally, as the field of natural language processing evolves rapidly, newer models and techniques may offer even more robust solutions in Arabic dialect identification and machine translation, necessitating ongoing research and adaptation.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, ElMoatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021a. 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.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021b. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022a. Nadi 2022: The third nuanced arabic dialect identification shared task. *WANLP 2022*, page 85.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022b. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. 2020. Arabic machine translation: A survey of the latest trends and challenges. *Computer Science Review*, 38:100305.

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 *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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 The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 31–42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rajaraman Kanagasabai, Saravanan Rajamanickam, Hariram Veeramani, Adam Westerski, and Kim Jung Jae. 2023. Semantists at ImageArg-2023: Exploring Cross-modal Contrastive and Ensemble Models for Multimodal Stance and Persuasiveness Classification. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.

Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich. 2023. Hybrid pipeline for building arabic tunisian dialect-standard arabic neural machine translation model from scratch. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–21.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic dialect identification using a parallel multidialectal corpus. In *International conference of the pacific association for computational linguistics*, pages 35–53. Springer.

Zinah J Mohammed Ameen, Abdulkareem Abdulrahman Kadhim, et al. 2023. Deep learning methods for arabic autoencoder speech recognition system for electro-larynx device. *Advances in Human-Computer Interaction*, 2023.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the 27th international conference on computational linguistics*, pages 1332–1344.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023a. Automated Citation Function Classification and Context Extraction in Astrophysics: Leveraging Paraphrasing and Question Answering. In *Proceedings of the second Workshop on Information Extraction from Scientific Publications*, Online. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023b. Enhancing ESG Impact Type Identification through Early Fusion and Multilingual Models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023c. KnowTellConvince at ArAIEval Shared Task: Disinformation and Persuasion Detection in Arabic using Similar and Contrastive Representation Alignment. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023d. LowResContextQA at Qur'an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection in Arabic. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023e. LowResourceNLU at BLP-2023 Task 1 2: Enhancing Sentiment Classification and Violence Incitement Detection in Bangla Through Aggregated Language Models. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023f. Temporally Dynamic Session-Keyword Aware Sequential Recommendation system. In *2023 International Conference on Data Mining Workshops (ICDMW)*.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.