Empirical Evaluation of Pre-trained Language Models for Summarizing Moroccan Darija News Articles

Azzedine Aftiss[1], Salima Lamsiyah[2], Christoph Schommer[2], Said Ouatik El Alaoui[1]

[1]Engineering Sciences Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco
[2] Department of Computer Science, FSTM, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Correspondence: azzedine.aftiss@uit.ac.ma

Abstract

Moroccan Dialect (MD), or "Darija," is a primary spoken variant of Arabic in Morocco, yet remains underrepresented in Natural Language Processing (NLP) research, particularly in tasks like summarization. Despite a growing volume of MD textual data online, there is a lack of robust resources and NLP models tailored to handle the unique linguistic challenges posed by MD. In response, we introduce **GOOD.MA_v2**, an expanded version of the **GOUD.MA** dataset, containing over 50k articles with their titles across 11 categories. This dataset provides a more comprehensive resource for developing summarization models. We evaluate the application of large language models (LLMs) for MD summarization, utilizing both fine-tuning and zero-shot prompting with encoder-decoder and causal LLMs, respectively. Our findings demonstrate that an expanded dataset improves summarization performance and highlights the capabilities of recent LLMs in handling MD text. We open-source our dataset, fine-tuned models, and all experimental code, establishing a foundation for future advancements in MD NLP. We release the code at https://github.com/AzzedineAftiss/Moroccan-Dialect-Summarization.

1 Introduction

Moroccan Dialect (MD), commonly known as "Darija," is the primary spoken variety of Arabic in Morocco, coexisting with Berber in some regions. Approximately 91% of Moroccans communicate in Darija [Ridouane et al., 2014]. With the rise of digital resources, MD textual data available online is rapidly growing [Labied and Belangour, 2021], creating a need for effective automatic summarization to help users extract key information efficiently. While extensive work has focused on widely spoken languages, such as English, limited research exists on MD [Tachicart and Bouzoubaa, 2022], especially in sequence-to-sequence (Seq2Seq) tasks like summarization.

Challenges in MD research include a lack of comprehensive corpora, limited linguistic resources, complex syntax that challenges NLP models, and unique vocabulary not present in standard Arabic lexicons. Although existing datasets, such as **GOUD.MA** [Issam and Mrini, 2021], offer foundational resources, the evolving nature of the MD necessitates additional, robust datasets. Motivated by these challenges, we introduce ***GOOD.MA_v2***, an expanded version of the **GOUD.MA** dataset, containing over 50,000 articles with titles across 11 categories, aiming to enhance model robustness for MD data. Additionally, we explore LLMs for summarizing MD text, analyzing various Seq2Seq models in addition to evaluating recent causal LLMs in a zero-shot prompting setting, thus contributing valuable insights into their performance on MD.

The emergence of LLMs has led to remarkable NLP advancements [Chang et al., 2024]. However, adapting these models for low-resource languages, including the MD, remains underexplored. Recent efforts have attempted to adapt Arabic-specific models (e.g., ArBERT [Antoun et al., 2020], DarijaBERT [Gaanoun et al., 2024], DziriBERT [Abdaoui et al., 2021]) and multilingual models (e.g., mBART [Liu, 2020], mT5 [Xue, 2020]) for dialectal Arabic [Khered et al., 2023, Nagoudi et al., 2021b, Smadi and Abandah, 2024, Fuad and Al-Yahya, 2022]. More recent models, such as GPT-4 [Achiam et al., 2023] and Llama 3 [Dubey et al., 2024], offer advanced NLP capabilities but have not yet been adapted to the MD.

In this paper, we conduct an empirical study of LLMs on **GOUD.MA_v2** specifically curated for abstractive summarization. We demonstrate that expanding the dataset with additional samples improves summarization performance. Our approach includes fine-tuning encoder-decoder models and adapting recent LLMs for zero-shot prompting, pro-

Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4), pages 77–85
January 20th, 2025.

viding comprehensive insights into the effectiveness of LLMs for MD text summarization. Our dataset, fine-tuned models, and all code used in our experiments are open-sourced.

The main contributions of this paper are as follows:

- We expand **GOUD.MA** to **GOOD.MA_v2**, comprising over 50,000 articles with their titles across 11 categories.
- We demonstrate that increasing the MD dataset improves model performance on summarization tasks in terms of ROUGE and BERTScore evaluation metrics.
- We empirically evaluate the effectiveness of various LLMs, including Seq2Seq and causal models, for MD summarization. To the best of our knowledge, this is the first study exploring the use of pre-trained language models for MD summarization.

The rest of this paper is organized as follows: Section 2 reviews related work on Arabic text summarization, with a focus on Arabic dialect summarization. Section 3 details the dataset collection process. Section 4 describes the experimental settings, while Section 5 presents the experimental results. Finally, Section 7 discusses the conclusions and limitations of this work.

2 Related Work

Arabic dialect processing has gained attention due to the linguistic diversity and widespread use of dialects in the Arabic-speaking world. Unlike Modern Standard Arabic (MSA), Arabic dialects, such as Moroccan Darija, exhibit unique lexical, syntactic, and phonological variations [ALFattah, 2024], presenting challenges for NLP tasks due to limited labeled data and resources. Early work in Arabic dialect NLP focused on tasks like classification [Maghfour and Elouardighi, 2018, Al-Walaie and Khan, 2017], identification [Elaraby and Abdul-Mageed, 2018, Zaidan and Callison-Burch, 2014, Salameh et al., 2018], and translation [Zbib et al., 2012, Harrat et al., 2019]. However, studies on dialectal summarization, particularly for MD, remain sparse. Issam and Mrini [2021] introduced one of the first MD summarization datasets, with articles paired with titles as reference summaries.

Arabic text summarization has advanced with methods like clustering, minimum redundancy–maximum relevance (mRMR), and graph-based approaches [Oufaida et al., 2014, Elbarougy et al., 2020]. Deep learning techniques, such as Seq2Seq architectures with LSTMs and attention mechanisms, have also been explored [Al-Maleh and Desouki, 2020]. Recent transformer-based models, such as AraBART [Eddine et al., 2022], AraT5 [Nagoudi et al., 2021a], and AraBERT [Antoun et al., 2020], pre-trained on large Arabic corpora, have shown strong performance in Arabic summarization tasks. Additionally, multilingual models like mBART [Liu, 2020] and mT5 [Xue, 2020], pre-trained on diverse language corpora, have demonstrated cross-lingual effectiveness, making them suitable for low-resource dialects, including Moroccan Darija. Recently, models like DarijaBERT [Gaanoun et al., 2024] and DziriBERT [Abdaoui et al., 2021] have been pre-trained on North African dialectal Arabic, specifically to address Moroccan and Algerian dialects. DarijaBERT, focused on Moroccan Darija, incorporates dialect-specific vocabulary, bridging the gap between MSA and regional dialects, thus enhancing contextual understanding compared to general Arabic models. The emergence of advanced LLMs, such as GPT-4o mini [Achiam et al., 2023], Llama 3 [Dubey et al., 2024], and Mistral NeMo [team, 2024] have brought attention to their capabilities in domain-specific tasks in zero-shot or few-shot settings. These models exhibit strong reasoning and text generation capabilities without fine-tuning task-specific data. However, applying them to dialectal summarization has limitations, as their training data generally lacks comprehensive coverage of specific dialects, such as Moroccan Darija. Fine-tuning remains essential to optimize performance for dialectal tasks. Our work builds upon these prior studies by applying and comparing Arabic-specific, multilingual, and causal LLMs for MD summarization using both zero-shot and fine-tuning methods. To our knowledge, this is the first work that evaluated LLMs for Moroccan Darija abstractive summarization, contributing valuable insights to NLP research for dialectal Arabic.

3 *GOOD.MA_v2*: A Newspaper Corpus for Moroccan Darija Summarization

3.1 Dataset Description

A primary challenge in NLP tasks for low-resource languages, such as Moroccan Darija, is the scarcity of high-quality datasets. To address this gap, Issam and Mrini [2021] recently introduced a benchmark dataset specifically for summarization, sourced from the GOUD.MA website[1].

GOUD.MA, a news website established by Ahmed Najim in 2011, is a primary source of Moroccan Darija text for summarization research. Articles on this platform are primarily in Arabic, with titles in Moroccan Darija and the body text in either Modern Standard Arabic (MSA) or a mix of MSA and Darija. The dataset is derived from GOUD.MA supports summarization tasks where the article text serves as input, and the title provides a concise target summary. Table 1 presents statistical details for the GOUD summarization datasets, including train, validation, and test splits.

Furthermore, the dataset covers various categories, such as الرئيسية (Main) and آش واقع (What's Happening), with each article assigned to a single category. This classification consists of a wide range of topics, from general news to specialized subjects like media, culture, and sports. Table 2 shows the distribution of articles across categories, including translations and article counts for each.

3.2 Data Collection

As previously mentioned, a major challenge in Moroccan Darija NLP tasks is the scarcity of large, annotated datasets. To address this, we expanded the **GOUD.MA** dataset by scraping additional articles and summaries from the GOUD.ma website. Expanding the dataset with additional text-summary pairs helps improve model performance by capturing a broader representation of linguistic patterns, expressions, and vocabulary unique to Moroccan Darija.

We utilized the Python libraries Scrapy[2] and Selenium[3] to automatically crawl the GOUD.ma website, collecting article titles, publication dates, content, and categories. This scraping process, conducted between 2022 and October 2024, took approximately four days and resulted in a total of 50,517 articles.

We cleaned and organized the collected data into CSV format for analysis, ensuring that each article entry includes metadata such as publication date, title, content, and category. Table 2 presents the distribution of articles across categories, while Table 1 provides statistics on article and title lengths. The expanded dataset aligns closely with previous datasets in terms of length distribution and is utilized to fine-tune models, enhancing performance on MD summarization tasks.

4 Experiment Settings

The main objectives of this study are twofold: (1) to expand the dataset to capture evolving dialectal variations and hence improve the performance of text summarization models, and (2) to evaluate the effectiveness of large language models in summarizing Moroccan Darija text. We conducted an empirical study to assess the performance of various LLMs, including pre-trained causal models and fine-tuned Seq2Seq models, on the ***GOOD.MA_v2*** dataset. In this section, we present the implementation details and a brief description of the models used for comparison.

4.1 Implementation Details

In this study, we applied three categories of models for MD text summarization: Arabic-specific models (AraBERT, DarijaBERT, DziriBERT, AraBART, and AraT5), a Multilingual Model (mBART), and causal large language models (GPT-4o mini, Llama 3, and Mistral NeMo). Following the approach of Rothe et al. [2020], which leverages pre-trained language models for abstractive summarization within a Seq2Seq framework, we fine-tuned DziriBERT, DarijaBERT, and AraBERT on our summarization dataset to capture linguistic nuances specific to Moroccan dialects, utilizing the strengths of encoder-based models. For AraT5, AraBART, and mBART, which already feature Seq2Seq architectures with both encoder and decoder components, we fine-tuned them directly for MD summarization. We used a merged dataset, combining the training set of **GOOD.MA** with the expanded **GOOD.MA_v2**, for training. For validation and testing, we used the original validation and test sets from **GOOD.MA**.

Each model was fine-tuned for 20 epochs with a batch size of 20, gradient accumulation set to 8,

[1] https://www.goud.ma/
[2] https://scrapy.org/
[3] https://selenium-python.readthedocs.io/

Dataset Split	Number of Articles	Avg. tokens per article	Avg. tokens per title
Train (GOUD.MA)	139,288	238.03	15.137
Validation (GOUD.MA)	9,497	238.54	15.14
Test (GOUD.MA)	9,497	238	15.20
Train (GOUD.MA_v2)	189,805	253.54	16.40

Table 1: Summary Statistics of the **GOUD.MA** and **GOUD.MA_v2** Dataset Splits. The "Number of Articles" column indicates the total count of articles in each split. The "Avg. tokens per article" represents the average number of tokens in the articles for each split. Finally, the "Avg. tokens per title" indicates the average number of tokens in the titles of the articles.

Category	Category Translation	Goud.MA Dataset (Number of Articles)	Goud.MA_v2 Dataset (Total Number of Articles)
الرئيسية	Main	104,724	132,392
آش واقع	What's happening	98,569	116,297
تبركيك	Gossip	16,867	17,827
كود سبور	Goud Sport	13,236	16,083
آراء	Opinions	8,239	8,585
ميديا وثقافة	Media and Culture	7,579	8,218
كودتيفي	Goud TV	6,966	7,043
الزين والحداكة	Beauty and Sharpness	5,223	5,297
جورنالات بلادي	National Newspapers	4,549	4,693
راس السوق	Market head	0	31
كود	Goud	1	4

Table 2: Distribution of Articles by Category. The "Category" column represents the name of the category, the "Category Translation" indicates the translation of the original category into English, the "**GOUD.MA** Dataset" column shows the number of articles from the old dataset, the "**GOUD.MA_v2** Dataset" column shows the number of articles from the expanded datasets.

weight decay of 0.01, and a learning rate of 2e-5. For text generation, we used beam search with a beam width of 5, a maximum input sequence length of 256, and a maximum target sequence length of 32. All models used in our study are available on Hugging Face [Wolf et al., 2020].

For the causal LLMs, we employed zero-shot prompting to adapt these models for MD summarization. Using the `unsloth`[4] library, which supports quantization techniques and parameter-efficient fine-tuning (PEFT) methods like Low-Rank Adaptation (LoRA), we optimized the LLMs to reduce computational resources and memory usage, facilitating deployment in resource-constrained environments. For GPT-4o mini, we used openAI API[5] to generate article summaries (titles). Figures 1, 2, and 3 illustrate the prompts used for the Llama 3, GPT-4 mini, and Mistral NeMo models, respectively.

```
You are a helpful AI assistant for generating a detailed title
that highlights the main ideas and topics of the article.
Please ensure the title is written in Arabic. Format the output
as follows:

### Text:
[Content of the Current Article]
### Title:
```

Figure 1: Prompt used for Llama 3 model.

4.2 Model Selection Criteria

As outlined earlier, the pre-trained models used in our study are grouped into three main categories: Arabic-specific models, multilingual mod-

[4] https://huggingface.co/unsloth
[5] https://platform.openai.com/

Generate a concise and coherent title in Arabic that highlights the main ideas and themes of the article.

Text:
[Content of the Current Article]
Title:

Figure 2: Prompt used for GPT-4o mini model.

Generate a title that accurately captures the main ideas and themes of the article.

Text:
[Content of the Current Article]
Title:

Figure 3: Prompt used for Mistral NeMo model.

els, and causal language models, which are briefly described below.

Arabic-Specific Models: These models are pre-trained on Modern Standard Arabic (MSA) and various Arabic dialects, making them well-suited for Moroccan Darija summarization. The Arabic-specific models used in this study include:

- **AraBERT** [Antoun et al., 2020]: A BERT-based transformer encoder pre-trained on large Arabic corpora, designed for masked language modeling across MSA and Arabic dialects.
- **DarijaBERT** [Gaanoun et al., 2024]: A BERT variant specifically pre-trained on Moroccan Darija, capturing its distinctive vocabulary and linguistic features.
- **DziriBERT** [Abdaoui et al., 2021]: A BERT-based model pre-trained on Algerian dialect, which shares linguistic similarities with Moroccan Darija, enhancing its relevance to this study.
- **AraBART** [Eddine et al., 2022]: An adaptation of the BART architecture, combining a bidirectional encoder and an autoregressive decoder, suited for sequence-to-sequence tasks like summarization.
- **AraT5** [Nagoudi et al., 2021a]: A variant of the text-to-text transformer (T5) pre-trained on MSA and various Arabic dialects, supporting a range of generative tasks.

Multilingual Models: Pre-trained on multiple languages, these models can handle diverse linguistic structures. We employed mBART [Liu, 2020], a Seq2Seq model pre-trained on numerous languages, including Arabic, using a denoising autoencoder to enhance performance across multilingual text generation tasks.

Causal Language Models: We evaluate three large language models — GPT-4o mini [Achiam et al., 2023], Llama 3 [Dubey et al., 2024], and Mistral NeMo [team, 2024]— in MD summarization using zero-shot prompting.

- **GPT-4o mini** [Achiam et al., 2023]: An autoregressive LLM with strong reasoning capabilities, supporting both text and vision inputs.
- **Llama 3** [Dubey et al., 2024]: A decoder-only transformer optimized for efficiency and robust across various language tasks.
- **Mistral NeMo** [team, 2024]: A LLM built on a transformer decoder architecture with a 128k-token context window, suitable for long-form summarization tasks.

5 Experimental Results

In this section, we present a comparative analysis of the pre-trained language models used on the ***GOOD.MA_v2*** dataset.

5.1 Evaluation Measures

To evaluate the quality of the generated summaries in this study, we used two automatic evaluation metrics: ROUGE [Lin, 2004] and BERTScore [Zhang et al., 2019]. ROUGE-1 measures the unigram overlap between the reference and generated summaries, while ROUGE-2 evaluates the bigram overlap. ROUGE-L calculates the longest common subsequence (LCS) between the reference and generated summaries, providing a measure of sequence similarity. BERTScore, on the other hand, measures similarity by comparing token pairs in the reference and generated summaries using contextual embeddings from the pre-trained BERT model.

6 Results

The results of our experiment are presented in Table 3 and Table 4. In Table 3, we report the performance of the BERT-based models (AraBERT, DarijaBERT, and DziriBERT), which we fine-tuned following the approach by Rothe et al.

[2020]. We adapted these models for sequence-to-sequence tasks within an encoder-decoder framework. We observe that fine-tuning the models on *GOOD.MA_v2* dataset improved their performance compared to previous results reported by Issam and Mrini [2021]. This improvement supports our hypothesis that a more diverse dataset enhances model generalization, enabling them to better handle the linguistic nuances of Moroccan Darija.

The other investigated models are reported in Table 4. Among the models, mBART achieved the highest performance, likely due to its Multilingual Denoising Pretraining on a large corpus covering 50 languages, which provides robust cross-lingual representations beneficial for Moroccan Darija summarization. AraBART and AraT5 also demonstrated competitive performance, leveraging their encoder-decoder architectures that were pre-trained end-to-end on Arabic text. This architecture effectively captures input context and generates abstract summaries, making it well-suited for MD summarization given the shared vocabulary between MSA and MD.

On the other hand, the extractive summarization baselines (Lead-2, TextRank, SumRank) and causal language models (GPT-4o mini, Llama 3, Mistral NeMo) achieved comparatively lower performance. The extractive baselines struggled to produce coherent summaries, as they simply concatenate sentences in an unsupervised manner, often leading to disconnected and inconsistent outputs. Similarly, causal LLMs like GPT-4o mini and Llama 3, which were employed using zero-shot prompting, did not perform as well as fine-tuned Arabic-specific and multilingual models. This is likely due to the lack of fine-tuning, which limits their adaptability to Moroccan Darija summarization tasks.

Moreover, relying exclusively on ROUGE and BERTScore metrics may not provide a comprehensive assessment of generative models. As noted by Nguyen et al. [2024], automatic evaluation metrics like ROUGE and BERTScore primarily measure n-gram overlap or embedding similarity, which may not fully capture the creative and contextually nuanced outputs generated by large language models (LLMs). For instance, LLMs like GPT-4o-mini and Llama 3 are capable of rephrasing summaries in ways that differ from the reference text but still convey the intended meaning. For the Mistral NeMo model, we observed difficulties in understanding Arabic texts and dialects, which led to hallucinations in the resulting summaries, Additionally, during the experiment, we found that Mistral NeMo sometimes generated the summary in French or Spanish. To mitigate this, we applied post-processing to translate these summaries into Arabic using the Google Translate API[6]. This translation step may have impacted Mistral's overall performance.

To illustrate the qualitative differences across models, we provide example summaries generated by each model alongside the reference summary in Table 5.

7 Conclusion

In this study, we conducted an empirical evaluation of various pre-trained language models for abstractive summarization of Moroccan Darija text, comparing the performance of Arabic-specific encoder-decoder models, multilingual models, and causal language models. Our findings demonstrate that the multilingual model mBART, which is pre-trained on a diverse set of languages, generally achieved superior performance compared to the other models. The zero-shot application of causal language models, including GPT-4o mini, Llama 3, and Mistral NeMo, showed potential; however, results indicated that fine-tuning would be necessary to achieve contextually accurate and fluent summaries in MD. Moreover, while automatic evaluation metrics like ROUGE and BERTScore provided useful quantitative insights, they may not fully reflect qualitative aspects such as readability, fluency, and consistency—attributes that are crucial in summarization tasks. Future research could further explore this area by fine-tuning causal language models for MD summarization and incorporating human evaluation to gain a more comprehensive understanding of summary quality.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Molham Al-Maleh and Said Desouki. 2020. Arabic

[6] https://github.com/ssut/py-googletrans

Models	GOOD.MA				GOOD.MA_v2			
	R-1	R-2	R-L	BERTScore	R-1	R-2	R-L	BERTScore
AraBERT [Antoun et al., 2020]	23.08	8.98	22.06	-	28.47	15.78	25.47	64.79
DarijaBERT [Gaanoun et al., 2024]	19.41	6.64	18.48	-	27.79	15.53	24.8	64.24
DziriBERT [Abdaoui et al., 2021]	17.98	5.83	17.22	-	24.16	12.42	21.89	63.11

Table 3: The performance result of the BERT-based models (AraBERT, DarijaBERT, and DziriBERT). The first set of results are from the models fine-tuned on the **GOOD.MA** dataset as reported by Issam and Mrini [2021], and the second set of results are from the models fine-tuned on the expanded **GOOD.MA_v2** dataset.

Models	R-1	R-2	R-L	BERTScore
Lead-2	17.19	8.67	15.14	56.66
TextRank	12.96	5.027	10.49	53.54
SumRank	12.47	4.724	10.44	54.86
AraBART [Eddine et al., 2022]	31.51	19.24	28.98	66.21
MBart [Liu, 2020]	**33.55**	**21.56**	**30.86**	**67.21**
AraT5 [Nagoudi et al., 2021a]	32.63	20.15	29.97	66.48
GPT-4o mini [Achiam et al., 2023]	18.05	7.75	18.9	60.07
Llama 3 [Dubey et al., 2024]	16.59	7.17	14.44	58.52
Mistral NeMo [team, 2024]	7.06	2.11	6.40	54.31

Table 4: Performance of Compared Models on the Test Set of the **GOOD.MA** Dataset. Fine-tuned models were trained on the expanded **GOOD.MA_v2** dataset.

text summarization using deep learning approach. *Journal of Big Data*, 7(1):109.

Mona Abdullah Al-Walaie and Muhammad Badruddin Khan. 2017. Arabic dialects classification using text mining techniques. In *2017 International Conference on Computer and Applications (ICCA)*, pages 325–329. IEEE.

Mohammed ALFattah. 2024. Morpho-lexical analysis of tehami arabic dialect. *Social Science and Humanities Journal (SSHJ)*, 8(06):4036–4076.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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*.

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274.

Reda Elbarougy, Gamal Behery, and Akram El Khatib. 2020. Extractive arabic text summarization using modified pagerank algorithm. *Egyptian informatics journal*, 21(2):73–81.

Ahlam Fuad and Maha Al-Yahya. 2022. Araconv: Developing an arabic task-oriented dialogue system using multi-lingual transformer model mt5. *Applied Sciences*, 12(4):1881.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2024. Darijabert: a step forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*, pages 1–13.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing & Management*, 56(2):262–273.

Abderrahmane Issam and Khalil Mrini. 2021. Goud.ma: a news article dataset for summarization in moroccan

Model	Generated Summary
Reference Summary	بعد الجفاف.. القرض الفلاحي خصص قروض إضافية بقيمة 1,5 مليار درهم للفلاحة.
GPT-4o mini Summary	مجموعة القرض الفلاحي تدعم الفلاحة بتدابير جديدة لمواجهة جفاف الموسم.
Llama 3 Summary	اتخذات مجموعة القرض الفلاحي، من بعد ما تشاورات مع وزارة الفلاحة والصيد البحري والتنمية القروية والمياه والغابات، مجموعة من التدابير اللي كتستهدف التخفيف من آثار ضعف التساق
Mistral Summary	أصدروا 1.5 مليار درهم للتخفيف من تأثير إكويا.
MBart	القرض الفلاحي خصص كريديات للفلاحة بقيمة 1,5 مليار درهم.
AraT5	القرض الفلاحي خصص غلاف مالي إضافي للقروض بقيمة 1,5 مليار درهم للتخفيف من آثار ضعف الشتا.
AraBART	القرض الفلاحي خصص 1,5 مليار درهم لإعادة جدولة مديونية الفلاحين.
AraBERT	القرض الفلاحي خصص 5, 1 مليار درهم لكريدي الفلاحي لفائدة الفلاحين الصغار.
DarijaBERT	القرض الفلاحي خصص 5, 2 مليار درهم للقروض الفلاحية بقيمة 5, 1 مليون درهم
DziriBERT	القرض الفلاحي خصص مليار درهم لتمويل ازمة كورونا

Table 5: The generated summaries of the models for a given article from the test set.

darija. In *3rd Workshop on African Natural Language Processing*.

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.

Maria Labied and Abdessamad Belangour. 2021. Moroccan dialect "darija" automatic speech recognition: a survey. In *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 208–213. IEEE.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Mohcine Maghfour and Abdeljalil Elouardighi. 2018. Standard and dialectal arabic text classification for sentiment analysis. In *Model and Data Engineering: 8th International Conference, MEDI 2018, Marrakesh, Morocco, October 24–26, 2018, Proceedings 8*, pages 282–291. Springer.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021a. Arat5: Text-to-text transformers for arabic language generation. *arXiv preprint arXiv:2109.12068*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021b. Investigating code-mixed modern standard arabic-egyptian to english machine translation. *arXiv preprint arXiv:2105.13573*.

Huyen Nguyen, Haihua Chen, Lavanya Pobbathi, and Junhua Ding. 2024. A comparative study of quality evaluation methods for text summarization. *arXiv preprint arXiv:2407.00747*.

Houda Oufaida, Omar Nouali, and Philippe Blache. 2014. Minimum redundancy and maximum relevance for single and multi-document arabic text summarization. *Journal of King Saud University-Computer and Information Sciences*, 26(4):450–461.

Tachicart Ridouane, Bouzoubaa Karim, and Jaafar Hamid. 2014. Building a moroccan dialect electronic dictionary (mded). In *5th International Conference on Arabic Language Processing*.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

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.

Malak Smadi and Gheith Abandah. 2024. Correcting auditory spelling mistakes in jordanian dialect using machine learning techniques. In *2024 15th International Conference on Information and Communication Systems (ICICS)*, pages 1–6. IEEE.

Ridouane Tachicart and Karim Bouzoubaa. 2022. Moroccan arabic vocabulary generation using a rule-based approach. *Journal of King Saud University-Computer and Information Sciences*, 34(10):8538–8548.

Mistral AI team. 2024. Mistral nemo. Accessed: November 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

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

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.