Universiteit Leiden

Sentiment Analysis of Moroccan Darija Arabizi: A Comparison of Pretrained Transformer Encoders and a Classical Baseline

Jakub Krajanowski (s4000544)

Supervisors:
Suzan Verberne & Max van Duijn

Bachelor Thesis– Data Science & Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 15, 2026

Abstract

The research on sentiment analysis of Arabic has focused on Modern Standard Arabic. Such approach leaves all the dialectal variations unattended. This thesis explores the possibilities of three-class sentiment classification for Moroccan Darija Arabizi, a spoken dialect, which suffers from low resources, no standard orthography and code switching. We compared five pretrained transformer encoders - two Darija-specific, one Arabic, and two multilingual - against a character n-gram SVM baseline. The dataset we used for fine-tuning was first derived from the Moroccan YouTube Corpus. Then it was manually re-labeled by native Darija speakers. The models are evaluated under a controlled protocol. Using stratified five-fold cross-validation over five seeds and three learning rates minimizes the probability of fortunate seeds. We chose the paired-bootstrap significance tests and evaluation metrics to take imbalance into account and allow final assessment. The encoders used for the experiment share the same architecture, therefore the performance differences are caused by pretraining and tokenizer fit, not model size. DarijaBERT-arabizi performs the best across all models, achieving a macro-F1 of ≈ 0.56. The 170M-parameter dialect-specific DarijaBERT-arabizi outperforms multilingual models like XLM-R with 278M parameters. The baseline performs comparably well to the transformers on accuracy, but falls behind on macro-F1 and chance-corrected agreement. The XLM-R is proven to collapse at higher learning rates. This work shows that specialization can outweigh scale in the case of small, low-resourced dialects and highlights the significance of investments into gathering resources for them.

Contents

1 Introduction

Sentiment analysis is one of the most crucial tasks of Natural Language Processing (NLP). Used in numerous processes, from product-review mining to tracking social media activity, it is utilized in all industries, from marketing to security. Bo Pang and Lillian Lee define sentiment analysis as: "the computational treatment of opinion, sentiment, and subjectivity in text" [32]. Today, it is mostly applied to user-generated, informal text. However, the progress in development of datasets, models, and benchmarks that drive it is uneven. Focusing mostly on the few high-resource languages, the development has left many widely spoken languages, and therefore people, behind.

A great example of this imbalance is the Arabic language. Concentration on formal written standards at the expense of spoken dialects is a general pattern in NLP, not unique to Arabic. Arabic illustrates it sharply: although it is one of the most widely spoken languages in the world, research on Arabic sentiment analysis has focused on Modern Standard Arabic (MSA), leaving the users of its many dialects under-served [6, 26, 3].

The main focus of this thesis is Moroccan Darija, an Arabic dialect used widely in Morocco. Darija, being a spoken dialect, has features like no standardized orthography, a very specific alphabet (Arabizi) and impact from other languages. This is illustrated in Table 1. These properties, together with the lack of useful training data, make sentiment analysis for Darija a challenging problem.

The typical approach to sentiment classification is fine-tuning a transformer encoder[14]. There are multiple models that could be applied to solve this problem, however the choice is far from obvious. Spanning from big multilingual encoders, through Arabic-focused models, to lightweight classical ones there are many options. This issue raises three practical questions: which encoder performs best on Moroccan Darija Arabizi, how robust the ranking is across runs, and whether the heavy models are worth their cost in that case. Prior work exists, but it is limited [4, 31]. What is missing is a rigorous comparison in a controlled environment.

Table 1: Example comments from the corpus, with English glosses and gold labels

Comment (Darija in Arabizi)	Idiomatic English gloss	Label
Je suis dgot	"I'm disgusted" (entirely French: *je suis dégoûté*).	−1
3sbatni had lmosi9a dyalk ikh 3la hala hram	"Your music got on my nerves — ugh, what a state, it's shameful."	−1
La 7awla wala 9owata ila bilah	"There is no power nor strength save in God" (set phrase voicing dismay).	−1
Ma 9dertch hta nesm3 sawto wla nchofo	"I couldn't even hear his voice or see him."	0
Ntlb mank diri vido 3la dak chaba7 dyl micheal Jackson	"I'm asking you to make a video about that Michael Jackson 'ghost'."	0
Congrats 1 million subscribers safaa	"Congrats on 1 million subscribers, Safaa" (entirely English).	+1
story time dyalk 14k..15k les vues 10k...6k 5M I love you	"Your story time — 14k..15k, the views 10k...6k, 5M — I love you."	+1

Comments are quoted verbatim from the annotated corpus; *Label* is the final adjudicated sentiment label in our dataset (−1 negative, 0 neutral, +1 positive). The examples show the three properties that make the task hard: phonemic digits standing in for Arabic consonants (*9* = qaf in *9dertch*, *3* = ayn in *3la*, *7* = haa in *7awla*), French/English code-switching (*je suis dégoûté*, *les vues*, *story time*), and the absence of a standard spelling, with one word written many ways (Table 2).

Table 2: Orthographic variation: one Darija word written many ways in the corpus

Word (meaning)	# spellings	Attested forms (with corpus frequency)
bzaf "a lot"	25	bzaf (30), bzaaf (8), bzaaaaf (7), bezaf (5), bezzaf (4), bzzaf (2), . . . , bzzz. . . zaf (1)
wllah "I swear"	9	wlah (26), walah (13), wllah (10), wallah (10), wellah (4), wlhh (2), . . .
wa3er "great/tough"	5	wa3ra (14), wa3er (8), wa3r (6), wa3rin (2), wa3rrr (1)

Counts are token frequencies over the 2,788-comment corpus. The variation comes from vowel choice, consonant doubling, and elongation: a single word for "a lot" appears in 25 different spellings, from *bzaf* up to *bzzz. . . zaf*. This same irregularity forces non-Arabizi tokenizers to split one word across many pieces (Table 4).

This thesis fills that gap. We evaluate five pretrained encoders and compare them against a baseline model. All of the models we have chosen span the spectrum explained above. Every model is trained and evaluated under an identical protocol, in a strictly specified environment and using concrete metrics for reproducibility.

Concretely, this thesis addresses the following research questions:

- **RQ1.** Which pretrained encoder performs best for three-class sentiment classification of Moroccan Darija Arabizi?
- **RQ2.** Is that ranking influenced by the choice of learning rate, random seed, and cross-validation fold?
- **RQ3.** Do the transformer encoders perform better than a classical char-n-gram baseline? Does the answer depend on whether performance is measured by accuracy or by metrics that take imbalance into account?
- **RQ4.** How does dialect-specific pretraining manifest - in tokenizer fit and in error patterns across Darija specific axes?

This thesis makes three contributions. First, we construct a three-class sentiment dataset for Moroccan Darija Arabizi by deriving an Arabizi-only subset of the Moroccan YouTube Corpus [22, 23] and re-annotating its 2,788 comments with negative, neutral, and positive labels, produced by two native speakers following written guidelines (Appendix A). Second, we provide a controlled comparison of five pretrained encoders and a char-n-gram SVM baseline under an identical protocol - stratified five-fold cross-validation, five seeds, and three learning rates - with imbalance-aware metrics and paired-bootstrap significance tests. Third, we show that dialect-specific pretraining outweighs scale on this task, trace this advantage to tokenizer fit and pretraining match, and document both the misleading behaviour of accuracy under class imbalance and the instability of XLM-R at higher learning rates.

Code and data availability. All code and the re-annotated dataset used in this thesis are publicly available at jkrajanowski/darija-arabizi-sentiment.

2 Background

Moroccan Darija is one of the dialects spoken daily by native Moroccans. It is the first language of most of the population, historically being only a spoken language, without a specified

written form [6, 26]. It consists of a Maghrebi-Arabic base, strongly influenced by Amazigh and borrowing vocabulary from languages like French, Spanish and English. The lexicon, morphology and phonology diverge from Modern Standard Arabic (MSA) [6, 26].

Modern Standard Arabic, despite being the formal written standard, is not spoken natively in Morocco, but rather learned in school. The language used online by the Moroccans is the dialect rather than the standard. Therefore, Arabic NLP resources transfer poorly to Darija, as they are built for MSA, which differs in morphology, vocabulary and spelling [6, 26].

The lack of standardized orthography in Moroccan Darija forces the speakers to improvise online using both Arabic script and Latin letters. This results in multiple variations of the same words with no canonical spelling. In this thesis we focus on Arabizi, the dominant alphabet used online and in messaging by Darija speakers. It is an informal romanization of the spoken language using the Latin alphabet and numerals for Arabic consonants with no Latin equivalent, for example, 3 = ayn. These digits are phonemic [13], not decorative, and they must be kept to ensure understandability. Darija, and therefore Arabizi, includes frequent French and English code-switching. Words from these languages are mixed into sentences, often affecting their sentiment.

Table 3: Arabizi digits and the Arabic consonants they encode

Digit	Arabic consonant	Approximate sound	Example from the corpus
2	hamza	glottal stop	*bi2dni* “by leave of”
3	ayn	voiced pharyngeal	*3la* “on/about”, *3lik* “to you”
5	kha	“ch” as in Bach	*5la3ni* “it scared me”
7	haa	breathy emphatic h	*7adi9a* “park”, *7dak* “next to you”
9	qaf	deep uvular k	*9dertch* “I couldn’t”, *mosi9a* “music”

The digits are phonemic, not decorative [13]: they stand for Arabic consonants that have no Latin letter and are chosen for visual resemblance. Example tokens are quoted verbatim from the corpus.

The specifics of Arabizi make it difficult to be processed by the subword tokenizers that transformer encoders rely on. As each encoder maps text onto a vocabulary of subwords (learned from pretraining [35, 25]), applying a model to a different script than the pre-training language leads to problems: tokenizers trained on MSA or multilingual text will break Arabizi words into multiple short pieces, separating the phonemic digits, as they were not exposed enough to Arabizi forms [34]. The over-fragmentation, caused by the lack of accurate subwords, results in the signals carried by each word being diluted and different spellings of each word being split differently. Tokenizers trained on Arabizi keep these specific forms in their original version, or split them accurately, without over-fragmentation. Table 4 shows this.

Table 4: Examples of tokenizer segmentation on two Arabizi comments

Input	Tokenizer	Segmentation
wa3er bezzaf hadchi	DarijaBERT-arabizi	`wa3er` | `bez` | `##zaf` | `hadchi` (4)
	MARBERTv2	`wa` | `##3` | `##er` | `be` | `##zz` | `##af` | `had` | `##ch` | `##i` (9)
	XLM-R	`_wa` | `3` | `er` | `_bez` | `za` | `f` | `_had` | `chi` (8)
ana 3ndi 7ess l9hwa	DarijaBERT-arabizi	`ana` | `3ndi` | `7es` | `##s` | `l9hwa` (5)
	MARBERTv2	`ana` | `3` | `##nd` | `##i` | `7` | `##ess` | `l` | `##9` | `##h` | `##wa` (10)
	XLM-R	`_ana` | `_3` | `ndi` | `_7` | `es` | `s` | `_l` | `9` | `h` | `wa` (10)

`##` marks a word-continuation piece and `_` a SentencePiece word boundary. The number in parentheses is the resulting token count. The DarijaBERT-arabizi tokenizer keeps Arabizi words and their phonemic digits together. Tokenizers trained on MSA and multilingual text fragment them and split the digits off.

3 Related Work

3.1 Sentiment analysis of Arabic and its dialects

Previous work on the topic of sentiment analysis on Arabic and its dialects shows a shift in the dominant approach. The latest papers focus on fine-tuned pretrained transformer encoders, the approach that shifted from classical pipelines based on TF-IDF features with Naive Bayes classifiers. The spoken dialects are under-resourced, as the resources and tools are concentrated on Modern Standard Arabic [6, 26, 3]. It is worth noting that the dialects are not minor variants. According to Badaro et al. the difference between dialects and MSA is comparable to the difference between Romance languages and Latin and the morphological tools created for MSA perform poorly on dialects [6]. Matrane et al. discuss the differences spanning phonological, morphological, lexical, and syntactic levels [26].

3.2 Sentiment resources for Moroccan Darija

As Moroccan Darija is a predominantly spoken dialect, the available written resources are limited. Therefore, data is often collected from social media, which makes it natural, and also valuable for sentiment analysis. The resource used in this thesis is the Moroccan YouTube Corpus (MYC) of Jbel et al. [22, 23]. It consists of 20,000 comments in mixed Arabizi and Arabic scripts scraped from Moroccan YouTube channels, labeled as positive or negative. Other available datasets differ mostly in origin of the content. The Moroccan Sentiment Twitter Dataset (MSTD) collects comments and posts from Twitter (currently X) [28] and the Moroccan Arabic Corpus (MAC) is built from Twitter posts [20]. Other collections exist, covering additional platforms and domains [18, 17, 9].

All of these datasets have similar traits: as the content comes from social media, the texts are short, informal, and orthographically irregular. French and English code-switching appears often across the data.

3.3 Pretrained encoders for Arabic and its dialects

The pretrained encoders differ in how much exposure to Arabic and dialects they have had in training. In multilingual models like mBERT or XLM-R trained on 104 and 100 languages respectively [14, 12], Arabic and its dialects are underrepresented, because of scale. Arabic-specific models target specific subgroups: The AraBERT is focused on Modern Standard Arabic

[5], whereas MARBERT and MARBERTv2 are trained on Arabic tweets that include dialectal content, but focused on the Arabic script [2]. CAMeLBERT explores pretraining variants – MSA, dialectal and classical – and their effects on performance [21]. The models closest to focusing on a singular dialect are DziriBERT[1], which targets Algerian, and TunBERT[27], targeting Tunisian. For Moroccan Darija, Gaanoun et al. released DarijaBERT in several variants, two of which we fine-tune and evaluate in this thesis: DarijaBERT-mix, pretrained on mixed Arabic- and Latin-script Darija, and DarijaBERT-arabizi, pretrained specifically on Latin-script Arabizi [19]. The progression along the axis of specificity is the span along which this thesis measures the effects of pretraining.

3.4 Comparative studies and evaluation of Darija sentiment

Previous work compares different options for sentiment analysis on Moroccan data. According to Amnay et al. contextual transformer embeddings perform better than feature-based techniques for Darija sentiment [4]. Recent work provides multi-dialect Arabic sentiment benchmarks that include Moroccan data [31]. Focusing on sentiment classifiers, community work on DarijaBERT has been published [30, 7]. However, the limitations noted previously can be found in these comparisons: they tend to fix a single learning configuration, on one evaluation split and report F1 scores or accuracy without significance testing.

4 Method

4.1 Dataset

The base dataset for our experiments is the Moroccan YouTube Corpus (MYC) [22, 23]. It consists of 20,000 real, public comments scraped from videos of Moroccan YouTube channels, written in Arabic script, in Arabizi, or in a mix of both. Each comment carries a binary sentiment label - positive or negative - assigned by the corpus authors. The comments span a wide spectrum of sentiment, from heavily negative (including racism, homophobia, and other forms of hate speech) to highly positive.

We do not reuse these labels. Our task is three-class classification, and a binary scheme cannot express the neutral class. We therefore discard the original labels and ask native Darija speakers to re-annotate an Arabizi-only subset of the corpus with negative, neutral, and positive labels. The selection of this subset and the annotation procedure are described in Section 4.2. One property of MYC is convenient for this re-annotation: the dataset contains only the comment text, without the video or surrounding discussion. Annotators therefore judge each comment in isolation - on exactly the information the classifiers later receive - so the labels match the input the models see.

4.2 Preprocessing and annotation

To mitigate bias caused by non-textual elements and minimize unwanted normalization of the data, the steps taken during preprocessing were deliberately minimal. We first preprocessed the dataset by removing emojis and non-text symbols and keeping only Arabizi comments, removing mixed and Arabic-script-only rows. As the original labels were only binary, to add the neutral class we removed the original labels and asked native speakers to re-annotate the subset. The preprocessing resulted in a subset of 4,700 comments and the first 3,000 of them were used for re-annotation.

Table 5: Class distribution of the final corpus (2,788 comments)

Class	Count	Share
Negative	381	13.7 %
Neutral	543	19.5 %
Positive	1,864	66.9 %
Total	**2,788**	**100 %**

The 2,788 comments include 2,546 agreed and single-annotator labels plus 242 adjudicated disagreements. Inter-annotator agreement on the 589 comments: $P_o = 0.589$, Cohen's $\kappa = 0.219$, quadratic-weighted $\kappa = 0.221$. The majority (always-positive) baseline accuracy is therefore 0.669.

The sample was split between two annotators, with a designed overlap of 600 comments labelled by both annotators. The split pattern for annotators A and B was as follows: 50 both / 200 A / 50 both / 200 B, repeating. The annotators are both female native Moroccans, 21 years old. One of them proposed the annotation guidelines with definitions for each sentiment label (See Appendix A) and then both of them were asked to follow the given definitions. In the end annotator A produced 1,680 unique labels and annotator B produced 1,697 unique labels, with 589 comments labelled by both of them. The ignored comments were not understandable, therefore they were removed from the dataset.

The annotators agreed on 347 out of 589 comments, which results in the following values: raw $P_o = 0.589$, Cohen's κ =0.219, quadratic-weighted $\kappa = 0.221$ [10, 11]. This outcome symbolizes a "Fair" agreement, taking into account that sentiment on short noisy text is subjective. It is worth noting that annotator A labelled 324 comments as negative in this subset, while annotator B chose the negative option in only 140 cases. Therefore, most disagreement exists on the negative–neutral boundary. Agreement at this level also matters for evaluation: human agreement acts as an approximate ceiling on the performance a model can reach on these labels, and we return to this point in the Discussion (Section 6). The 242 comments, on which the annotators did not agree, were adjudicated by randomly giving them one of the labels assigned by the annotators on a fixed seed. This approach was suggested by the annotators themselves, as many of these comments were sayings and religious phrases, which could be understood very differently based on the context. In the end, the dataset consists of 2,788 annotated comments, with the class distribution visible in Table 5 below.

4.3 Models compared

We compare five pretrained transformer encoders, specifically chosen to span the pretraining languages and alphabets most relevant to Arabizi. They are compared against each other and a SVM-based baseline with n-gram features. The encoders used are:

1. DarijaBERT-arabizi - pretrained on Moroccan Darija written in Arabizi specifically [19]
2. DarijaBERT-mix - pretrained on Moroccan Darija in mixed Arabic and Latin script [19]
3. MARBERTv2 - Arabic, dialect-aware encoder, pretrained on around one billion Arabic tweets written in Arabic script [2]
4. mBERT - multilingual BERT, pretrained on Wikipedia pages in 104 languages [14]
5. XLM-R - pretrained on 100 languages from the CC-100 web corpus [12]

Table 6: Models Comparison

Model	Params (M)	Pre-training data	Script	Vocab	Subw./word
DarijaBERT-arabizi	170.5	Moroccan Darija	Latin (Arabizi)	110,000	1.44
DarijaBERT-mix	208.9	Moroccan Darija	mixed	160,000	1.51
MARBERTv2	162.8	Arabic (∼1B tweets)	mostly Arabic	100,000	2.63
mBERT	177.9	104 langs (Wikipedia)	mixed	119,547	2.32
XLM-R	278.0	100 langs (CC-100)	mixed	250,002	2.16
char-n-gram SVC	—	— (TF-IDF char 3–5-grams)	—	—	—

All five transformers share the same ≈86 M-parameter 12-layer/768-hidden encoder and differ in their pretrained weights and in embedding/vocabulary size. Subwords-per-word is measured over the whole 2,788-comment corpus.

Their pretraining data, script, and vocabulary are summarised in Table 6.

The choice of encoders allows a controlled comparison. Every encoder's architecture is similar: a 12-layer, 768-hidden, 12-head Transformer with a 3,072-dimensional feed-forward size and ≈86 M-parameter base configuration. The only differences between these encoders are their pretrained weights, because of the different corpora and the size of their vocabulary. Because of the architectures being similar, the total parameter count depends almost entirely on vocabulary size. For example, the dialect-specific DarijaBERT-arabizi with 170.5M parameters is smaller than the multilingual XLM-R. Therefore, any performance differences between them cannot be caused by a larger or deeper network, which allows the isolated comparison of pretraining data and tokenizers and their fit to Arabizi.

The baseline used for the experiments uses no pretrained language model. It represents each comment as a set of character n-grams of length 3 to 5, weighted by TF-IDF, to make the common sequences weigh less. Each n-gram has to appear in at least two comments to be kept. They are then fed to a linear support vector machine (SVM) to classify the comment, taking into account the class weight balance. We used the scikit-learn implementation [33]. The dependency on characters, not full words, makes the model handle multiple spellings of the same word in Arabizi. This baseline is a light reference point to judge whether the much heavier transformer models are worth their cost.

4.4 Experimental setup

Data splitting & cross-validation To compare outcomes for the whole dataset we decided to use stratified 5-fold Cross Validation. As the dataset is imbalanced, it is stratified on the 3-class label. Therefore, each fold preserves the 13.7/19.5/66.9% class ratio, which is important in that heavily imbalanced set. Within each fold, 70% is used for training, 10% as a validation set for early stopping and model selection, and the remaining 20% as an isolated test set on which all reported metrics are computed.

Seeds & Repetition The fine-tuning was performed on five different seeds, fixed and listed for reproducibility: 13, 42, 2024, 7, 101. To maximize reproducibility, the seeds control fold assignment and training stochasticity, including data order, dropout and initialization of the classifier head.

Within each seed, the corpus is partitioned into folds, so that each comment appears in exactly one test set. Therefore, the metrics for a given seed are calculated over the whole dataset,

not a single split. The mean and standard deviation of per-seed scores are then reported, which highlights differences between seeds.

Each (model,learning rate) pair runs 25 times, once per each seed-fold pair. Therefore, in total there are 375 runs (5 transformers, 5 seeds, 5 folds and 3 different learning rates for each) and 25 baseline runs, as it is not dependent on learning rate.

4.5 Evaluation and analysis

4.5.1 Metrics

We use macro-F1 as the headline metric. It averages per-class F1 scores with equal weight, which makes it impossible for one class to dominate the score. As the dataset is imbalanced, it is a viable metric to compare models and perform as the early-stopping criterion, not being affected by the imbalance. For the evaluation scores, the selected learning rate for each model is the one maximizing macro-F1.

We also report accuracy, but not use it as a headline, because it is very susceptible to the imbalance. The always-positive baseline of 0.669 would make the outcomes biased if other metrics were ignored.

We additionally report Per-Class F1 to measure where the models differ, as the headline gap will be expected to result from minority classes.

We report Cohen's κ and quadratic-weighted κ [10, 11] as a secondary check. Here κ measures the chance-corrected agreement between the model's predictions and the gold labels; not the inter-annotator agreement of Section 4.2. Because the accuracy is inflated by a dominant positive class, the chance-corrected agreement allows us to discount the agreement expected from the class proportions. The quadratic-weighted kappa is utilized to punish the positive-negative reversal more than negative-neutral or neutral-positive confusion. Therefore, the sentiment ordinality is taken into account and models that make mild mistakes score higher than models that confuse polar opposites.

Each metric is computed per seed on the whole dataset. Then, they are reported as mean and standard deviation over all five seeds.

4.5.2 Significance testing

The differences between models could be caused by a particular test split. Therefore, to ensure reliability, a paired bootstrap is applied on macro-F1 [24, 8, 16]. Each comparison consists of 1,000 resamples of the comment and both the resampled subset and learning rate are the same for each compared model. Both models are then scored and the difference in macro-F1 is recorded. The two-sided p-value references the proportion in which the sign of the difference reverses. The tests were run for each (learning rate, seed) pair and then summarized by the largest p-value across the five seeds.

Only if a difference holds for every seed, it can be treated as significant. We report a difference as significant when $p_{\max} < 0.05$, borderline when $0.05 \leq p_{\max} < 0.10$, and not significant otherwise.

4.5.3 Error analysis

To define where the models perform best and worst, the performance is broken down into three axes chosen specifically for Moroccan Darija Arabizi. The predictions are pooled from all five seeds (13,940 predictions in total). The axes are as follows:

- Code-Switching: a comment is flagged as code-switched if any of its tokens is found in a small hand-curated list of French and English words observed in the data
- phonemic-Digit density: the part of characters in a comment that are Arabizi phonemic digits, labelled as none, low (≤ 0.05), and high (> 0.05).
- Text Length: the amount of tokens, separated by whitespaces, labelled as very short (≤ 3), short (4–8), medium (9–20), and long (> 20)

Similar to before, we report macro-F1 and accuracy for each slice. Error analysis allows comparison between specific subgroups.

4.5.4 Confusion Analysis

The row-normalised confusion matrices are inspected, pooled over all seeds and folds. Each row in the matrices corresponds to a true class and is scaled to sum to one. Therefore, diagonal entries represent per-class recall. The matrices make a model's collapse onto a majority class visible, as they reveal which classes are confused with each other (highlighting whether the errors fall on adjacent classes or polar opposites).

5 Results

5.1 Overall ranking

In the overall summary, DarijaBERT-arabizi performs best at every learning rate. With macro-F1 staying consistently in the 0.553–0.560 range across all learning rates range it outperforms every other model (Table 7). Figure 1 shows that it also performs better than the other models in terms of Cohen's kappa. The best performance was achieved by DarijaBERT-arabizi at 1×10^{-5}, with macro-F1 0.560 and κ 0.366, while the highest accuracy of 0.658 was reached by this model at learning rate of 2×10^{-5}. These values are the best across every (model, learning-rate) pair on all three variables. It is worth noting that the error bars are small, as the macro-F1 presented standard deviation of ≤ 0.009 across seeds. Therefore, the lead represents a stable best performance, rather than a fortunate seed.

DarijaBERT-mix consistently performs as second best. With the macro-F1 of around 0.53–0.54 at every rate, the two DarijaBERT variations are clearly separated from the other models. MARBERTv2, mBERT, and the baseline are clustered together, all between 0.46 and 0.48 macro-F1. The worst performing model is XLM-R with macro-F1 of 0.38–0.43 across seeds. The difference between DarijaBERT-arabizi and the best performing non-DarijaBERT model is 0.07–0.09 dependent on the learning rate and it rises to 0.13 over XLM-R. Cohen's kappa, presented on the right panel of Figure 1 orders the models in the same way. Therefore, the order is not an artefact of the chosen metric.

It is worth noting that accuracy does not follow the order described above. The baseline model reaches 0.649 accuracy, which is the second highest score, outperformed only by the DarijaBERT-arabizi. However, its macro-F1 is only 0.471. Therefore, accuracy and macro-F1 show very different outcomes for the baseline. That behavior is traced to the minority classes in Section 5.2 and is analyzed as the imbalance trap in the Discussion (Section 6, Figure 8).

Table 7: Model × learning-rate leaderboard, sorted by macro-F1.

Model	LR	Accuracy	Macro-F1	Cohen κ	Quad. κ
DarijaBERT-arabizi	1×10^{-5}	0.654±0.009	**0.560**±0.008	**0.366**±0.009	**0.419**±0.013
DarijaBERT-arabizi	2×10^{-5}	**0.658**±0.012	0.555±0.009	0.359±0.012	0.410±0.015
DarijaBERT-arabizi	3×10^{-5}	0.652±0.009	0.553±0.005	0.354±0.006	0.400±0.005
DarijaBERT-mix	2×10^{-5}	0.626±0.029	0.543±0.020	0.341±0.029	0.388±0.030
DarijaBERT-mix	3×10^{-5}	0.622±0.019	0.533±0.015	0.322±0.028	0.373±0.036
DarijaBERT-mix	1×10^{-5}	0.615±0.006	0.532±0.007	0.325±0.009	0.362±0.014
MARBERTv2	2×10^{-5}	0.590±0.012	0.481±0.008	0.247±0.012	0.247±0.018
MARBERTv2	3×10^{-5}	0.606±0.009	0.478±0.008	0.241±0.014	0.247±0.022
char-n-gram SVC	—	0.649±0.003	0.471±0.006	0.228±0.010	0.253±0.016
mBERT	3×10^{-5}	0.603±0.011	0.466±0.009	0.221±0.013	0.227±0.012
mBERT	2×10^{-5}	0.591±0.014	0.458±0.011	0.211±0.016	0.215±0.018
MARBERTv2	1×10^{-5}	0.553±0.023	0.456±0.011	0.210±0.018	0.206±0.024
mBERT	1×10^{-5}	0.559±0.011	0.453±0.005	0.202±0.008	0.198±0.015
XLM-R	1×10^{-5}	0.548±0.029	0.428±0.017	0.156±0.027	0.128±0.031
XLM-R	2×10^{-5}	0.532±0.018	0.411±0.015	0.126±0.021	0.089±0.021
XLM-R	3×10^{-5}	0.525±0.040	0.380±0.019	0.081±0.029	0.049±0.028

Values are mean ± std over the 5 seeds (each seed scored on the full corpus, the union of its 5 folds). Best value in each column is bolded. The char-n-gram baseline is learning-rate independent (identical rows). Quad. κ is quadratic-weighted Cohen's κ.

Table 8: Per-class F1 at each model's best learning rate

Model	LR	$F1_{neg}$	$F1_{neu}$	$F1_{pos}$	Macro-F1
DarijaBERT-arabizi	1×10^{-5}	0.445±0.014	0.458±0.008	0.777±0.006	0.560±0.008
DarijaBERT-mix	2×10^{-5}	0.430±0.016	0.451±0.017	0.749±0.031	0.543±0.020
MARBERTv2	2×10^{-5}	0.316±0.012	0.399±0.011	0.729±0.011	0.481±0.008
mBERT	3×10^{-5}	0.286±0.013	0.365±0.014	0.746±0.009	0.466±0.009
XLM-R	1×10^{-5}	0.245±0.012	0.350±0.022	0.690±0.031	0.428±0.017
char-n-gram SVC	—	0.292±0.018	0.338±0.015	0.784±0.003	0.471±0.006

Best learning rate selected by macro-F1. Values are mean ± std over the 5 seeds (each seed scored on the full corpus, the union of its 5 folds). Performance separates on the minority classes.

5.2 Per-class performance

Testing per-class shows that the differences in overall ranking come almost entirely from the minority classes. Per-class F1 is reported in Table 8 for each model's best learning rate. Every model performs well on the positive class, with F1 spanning 0.69–0.78. However, for negative and neutral classes, the F1 scores are lower and spread widely (negative F1 from 0.245 to 0.445, neutral from 0.338 to 0.458). Macro-F1 weights these three classes equally, therefore, taking into account that the positive scores are so similar, the overall ranking in Section 5.1 is driven by how well each model handles the minority classes. DarijaBERT-arabizi performs best on both the neutral and negative class, achieving F1 scores of 0.445 (negative) and 0.458 (neutral). Figure 2 shows the discussed pattern at 2×10^{-5} learning rate.

The dataset imbalance is visible in the performance of the char-n-gram baseline. It achieves the highest positive F1 score of all models (0.784, slightly outperforming DarijaBERT-arabizi's at 0.777). However, its negative and neutral F1 scores drop to 0.292 and 0.338 accordingly.

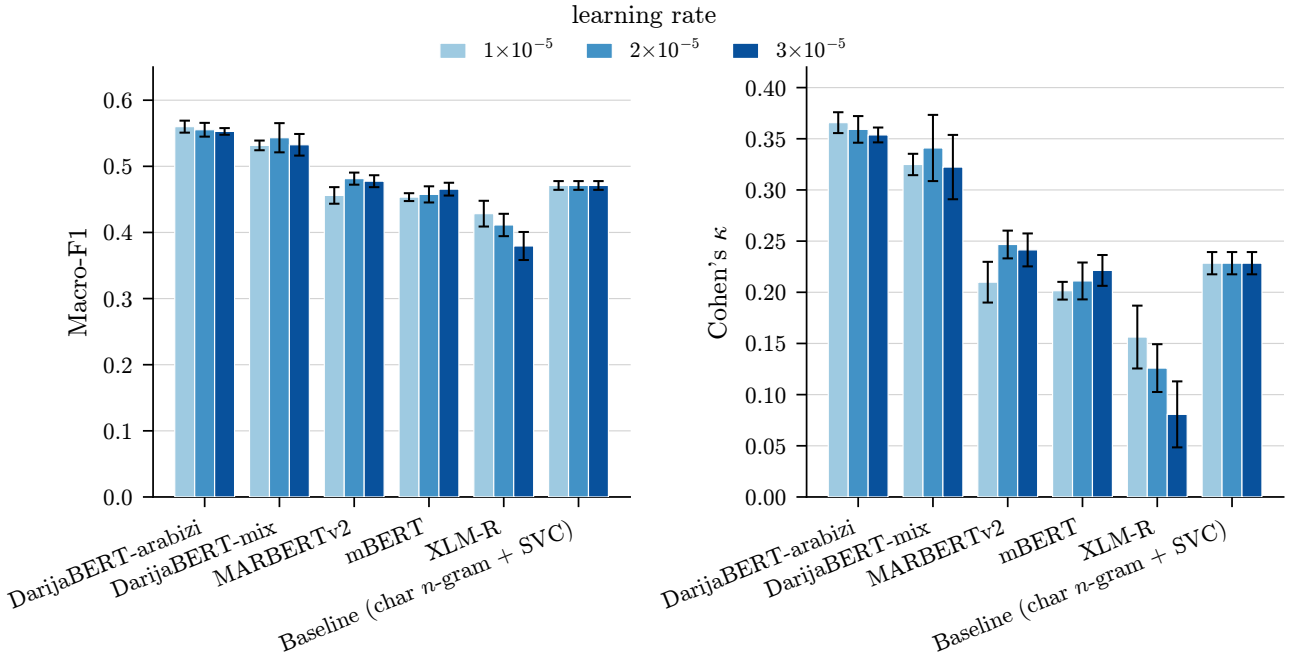

Figure 1: Headline results. Macro-F1 (left) and Cohen's κ (right) for each model at the three swept learning rates. Points are means over the five seeds and error bars span ± 1 standard deviation across seeds, each seed scored on the full corpus (the union of its five folds). DarijaBERT-arabizi leads at every learning rate on both metrics.

Therefore, its accuracy reported in Section 5.1 comes almost entirely from the majority class. The contrast between the overall best-performing DarijaBERT-arabizi and the baseline is worth highlighting. They are within 0.007 on positive score, but DarijaBERT-arabizi leads by 0.15 on negative and 0.12 on neutral, which causes its 0.089 macro-F1 advantage. This imbalance effect is analysed in the Discussion (Section 6, Figure 8).

5.3 Learning-rate sensitivity and the XLM-R collapse

For most of the models the learning rate does not have a strong effect on macro-F1. Table 9 reports macro-F1 at each rate and the per-model spread (max−min over the three rates). It is noticeable that every model except XLM-R varies by at most 0.025 (baseline is learning rate independent, because of its construction). The DarijaBERT-arabizi, with the highest scores, has a spread of only 0.007. Therefore, its lead does not depend on tuning the learning rate (Figure 3). The only other noticeable difference is shown for MARBERTv2. It under-fits at 1×10^{-5} (0.456), but then settles near 0.48.

XLM-R is the exception. As the learning rate rises, its macro-F1 falls from 0.428 ($\kappa = 0.156$) at 1×10^{-5} to 0.380 ($\kappa = 0.081$) at 3×10^{-5}, resulting in a spread of 0.049, significantly higher than any other in the table. That is caused by a single-class collapse. At 3×10^{-5} several runs started to over-predict one class. It is worth noting that the over-predicted class varies from run to run. Four runs collapsed to the positive class, achieving the accuracy of 0.67 (always positive) and three more collapsed to negative or neutral, dropping to 0.33 accuracy. This instability causes XLM-R's mean accuracy at 3×10^{-5} (0.525) to be below the 0.669 always-positive baseline. No other model shows similar behavior, MARBERTv2 having only one collapsed run and the remaining encoders none.

That outcome demonstrates the viability of a chance-corrected metric being kept alongside

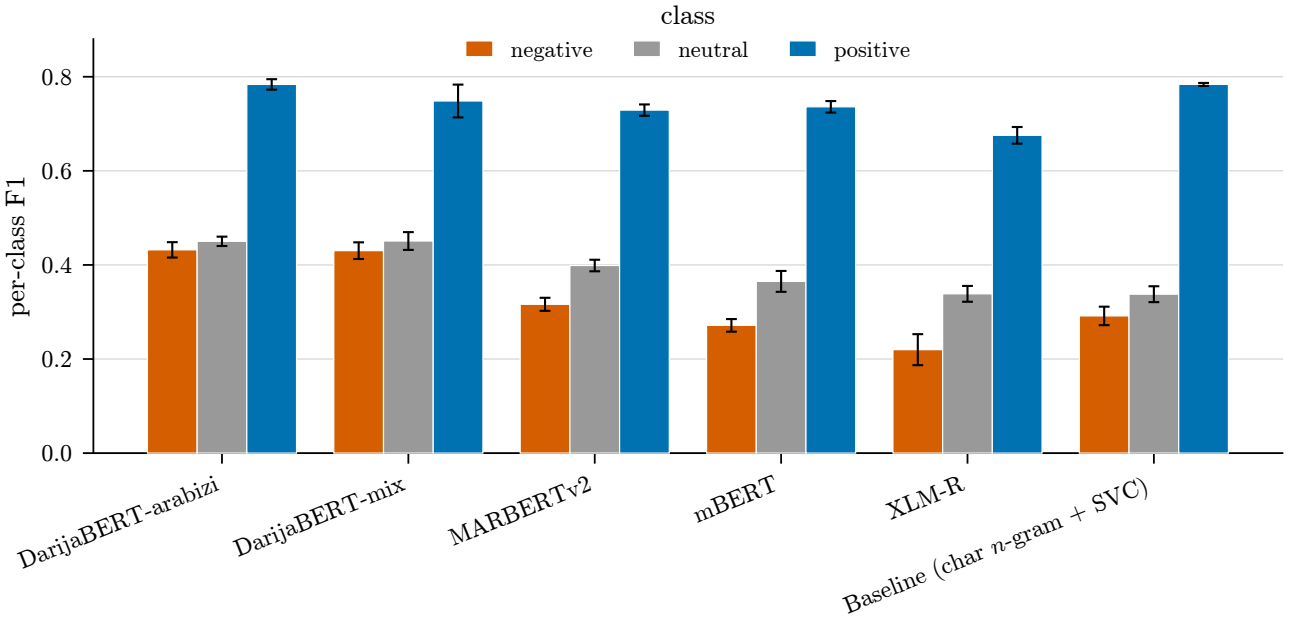

Figure 2: Per-class F1 at lr= 2×10^{-5} (± 1 s.d. across seeds). All models perform well on the majority positive class. Performance separates on the minority negative and neutral classes.

Table 9: Learning-rate sensitivity of macro-F1

Model	lr=10^{-5}	lr=2×10^{-5}	lr=3×10^{-5}	Spread
DarijaBERT-arabizi	0.560±0.009	0.555±0.010	0.553±0.005	0.007
DarijaBERT-mix	0.532±0.007	0.543±0.022	0.533±0.016	0.012
MARBERTv2	0.456±0.013	0.481±0.009	0.478±0.009	0.025
mBERT	0.453±0.006	0.458±0.011	0.466±0.010	0.012
XLM-R	0.428±0.019	0.411±0.017	0.380±0.021	0.049
char-n-gram SVC	0.471±0.007	0.471±0.007	0.471±0.007	0.000

Macro-F1 as mean $\pm$ std over the 5 seeds (each seed scored on the full corpus, the union of its 5 folds). Spread is max $-$ min across the three learning rates. XLM-R is the only model whose performance degrades with the learning rate. The char-n-gram baseline is learning-rate independent by construction.

accuracy. In the all-positive runs, XLM-R achieves 0.67 accuracy, but its kappa is very low. Instabilities of that kind are known in cases of fine-tuning large encoders on small, imbalanced datasets[15, 36].

5.4 Confusion analysis

The confusion matrices shown in Figure 4 are pooled over all seeds and folds, therefore each diagonal entry is a per-class recall. The recovery rates for each class correspond with their frequencies. The positive class has the highest recall 0.64–0.83, negative class has the lowest and neutral lies in between. The best performing model is DarijaBERT-arabizi with recall scores of 0.49 (negative), 0.49 (neutral), and 0.74 (positive). For the other models, negative recall falls to 0.25 for the baseline. Therefore, the hardest class is negative.

The error types depend on the model. Both DarijaBERT variations make negative errors almost evenly between the neutral and positive class. The other encoders and the baseline tend to make more negative errors in the majority positive class. Negative→positive rises to

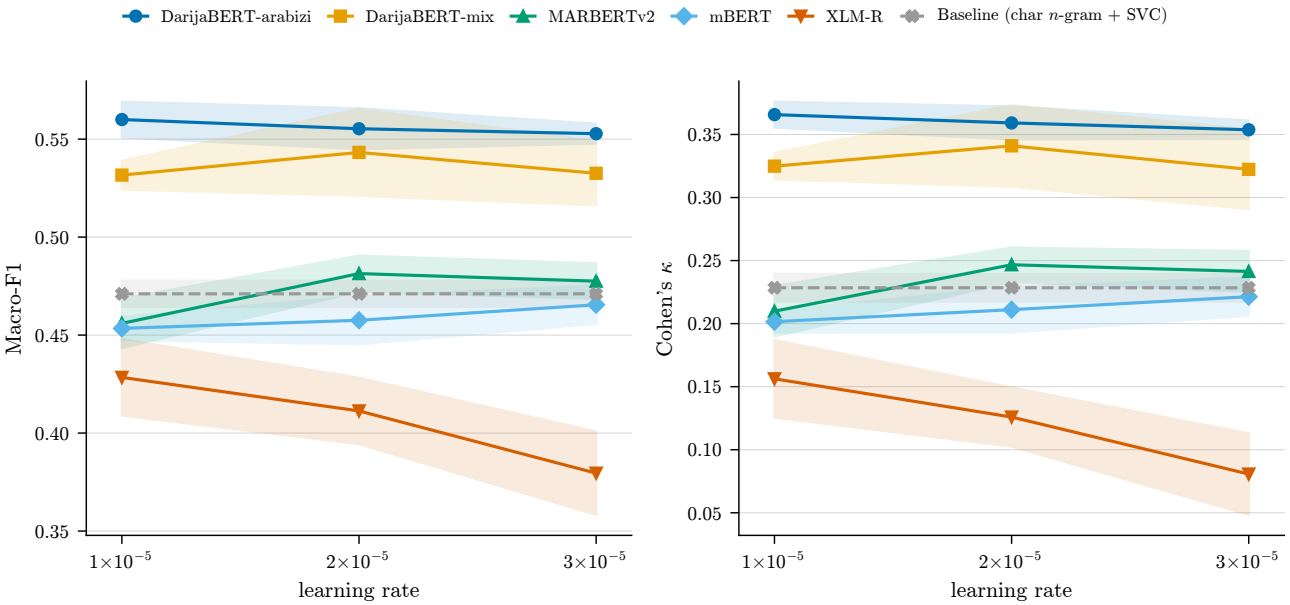

Figure 3: Learning-rate sensitivity of macro-F1 (left) and Cohen's κ (right). Shaded bands are ± 1 s.d. across seeds. DarijaBERT-arabizi is flat and highest. XLM-R degrades with higher LR and class-collapses at 3×10^{-5}.

0.54 for XLM-R and 0.58 for the baseline. Although the reverse error is rare, as positives are almost never labeled negative (0.06 - 0.19). Therefore, for the stronger models the errors appear more frequently between adjacent classes than between negative and positive. The quadratic-weighted kappa penalises the positive-negative mistakes more than confusions between adjacent classes. That can be shown for two edge cases: for DarijaBERT-arabizi unweighted kappa is 0.359 and quadratic-weighted kappa is 0.410 and for XLM-R unweighted kappa is 0.126 and quadratic-weighted kappa is 0.089 at 2×10^{-5} (Table 7).

5.5 Significance

Table 10 reports DarijaBERT-arabizi against every other model. It is noticeable that the difference between DarijaBERT-arabizi and all non-DarijaBERT models is significant for every learning rate. It achieves $p_{\max} < 0.001$ and margins between +0.07 and +0.17 macro-F1. However, the outcome is different for DarijaBERT-mix, which is an exception. With outcomes as follows: $p_{\max} = 0.700$ at 2×10^{-5}, $p_{\max} = 0.698$ at 3×10^{-5}, and borderline at 1×10^{-5} ($p_{\max} = 0.054$), they perform very closely.

The full pairwise matrix confirms the earlier outcomes. The DarijaBERT models beat every non-DarijaBERT model significantly. The remaining models form a cluster, where MARBERTv2, mBERT, and the char-n-gram baseline are mutually non-significant at every learning rate (well above the 0.05 threshold, with smallest $p_{\max} \approx 0.39$). XLM-R performs the worst, staying within the cluster at 1×10^{-5}, but falling below all the other models at learning rates 2×10^{-5} and 3×10^{-5}.

5.6 Error analysis along Darija-specific axes

Table 11 breaks DarijaBERT-arabizi's predictions (at 2×10^{-5}, pooled over the five seeds) down along three axes specific to Darija Arabizi. Figure 6 shows the same breakdown for every model.

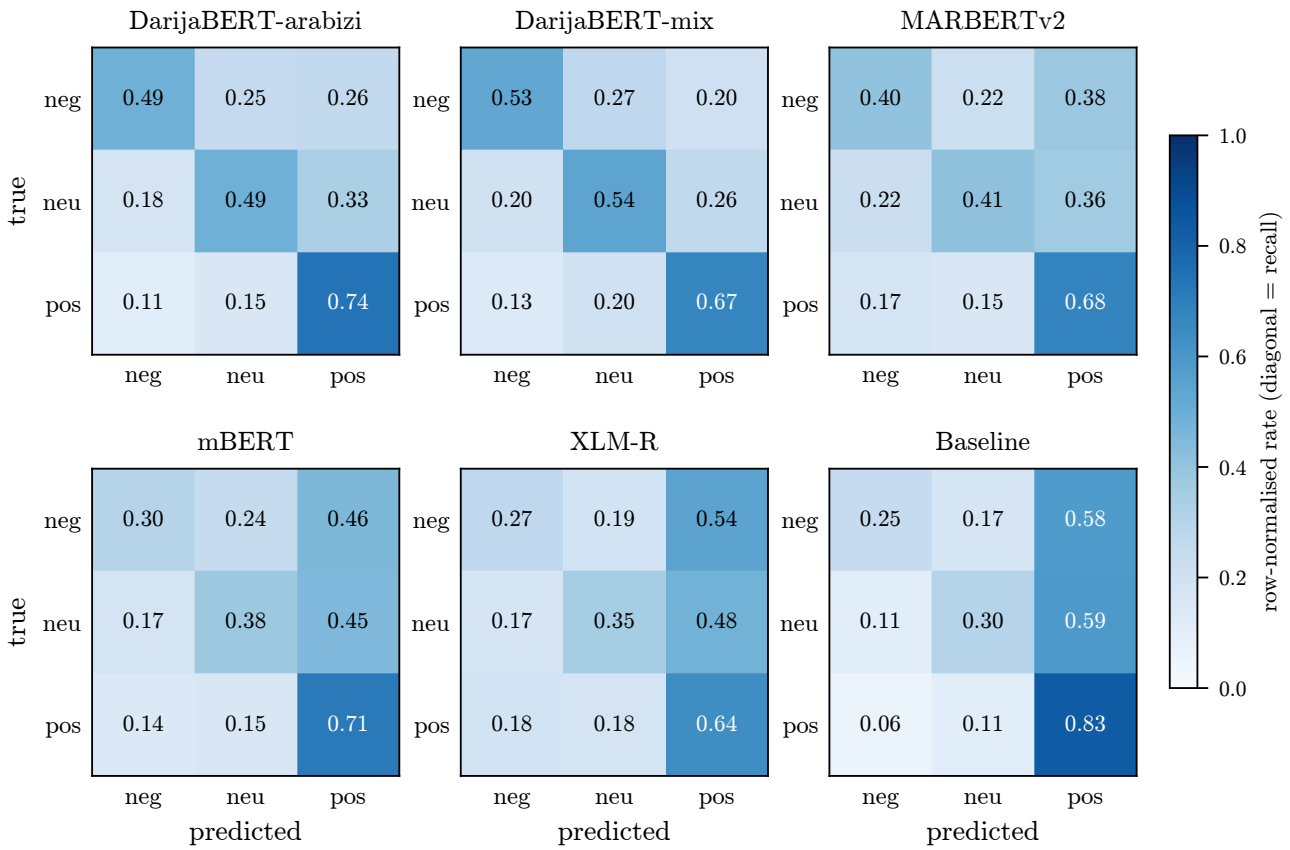

Figure 4: Row-normalised confusion matrices at lr= 2×10^{-5} (diagonal = recall), pooled over all seeds and folds. Negative is the hardest class for every model and is increasingly confused for positive by the weaker encoders, while the reverse is rare. XLM-R's full single-class collapse occurs only at 3×10^{-5} (Section 5.3).

5.6.1 Code-switching

Comments that included French or English words were classified more accurately than comments not containing code-switching, with accuracy scores of 0.831 and 0.640. However, the macro-F1 is almost the same for each subset (no code-switching: 0.552 and code-switching: 0.536). That suggests that the difference in accuracy comes from class composition. That is true, as the code-switched slice is overwhelmingly positive at about 86%, against 65% elsewhere. The accuracy of a model rises, by simply defaulting to positive class. Therefore, the gain on accuracy is an artifact of class composition and not a result of better understanding.

5.6.2 Phonemic-digit density

The macro-F1 scores decline as the share of Arabizi digits rises. Example values are 0.561 (no digits) through 0.541 (low) to 0.514 (high) for DarijaBERT-arabizi. The decline happens in every model, but the Darija encoders remain the best at every level. DarijaBERT-mix has the least differences, varying by only 0.03 across the three buckets. That is consistent with the Arabizi tokenizers keeping phonemic digits intact, not separating them (Section 5.7).

Table 10: Pairwise paired-bootstrap significance of DarijaBERT-arabizi versus every other model

LR	vs. model	Δmacro-F1	$p_{\max}$	verdict
1×10^{-5}	DarijaBERT-mix	$+0.028_{\pm 0.009}$	0.054	borderline
	MARBERTv2	$+0.104_{\pm 0.016}$	0.000	sig.
	mBERT	$+0.107_{\pm 0.006}$	0.000	sig.
	XLM-R	$+0.132_{\pm 0.014}$	0.000	sig.
	char-n-gram SVC	$+0.089_{\pm 0.013}$	0.000	sig.
2×10^{-5}	DarijaBERT-mix	$+0.012_{\pm 0.015}$	0.700	n.s.
	MARBERTv2	$+0.074_{\pm 0.007}$	0.000	sig.
	mBERT	$+0.098_{\pm 0.006}$	0.000	sig.
	XLM-R	$+0.144_{\pm 0.012}$	0.000	sig.
	char-n-gram SVC	$+0.084_{\pm 0.011}$	0.000	sig.
3×10^{-5}	DarijaBERT-mix	$+0.020_{\pm 0.017}$	0.698	n.s.
	MARBERTv2	$+0.075_{\pm 0.006}$	0.000	sig.
	mBERT	$+0.087_{\pm 0.012}$	0.000	sig.
	XLM-R	$+0.173_{\pm 0.019}$	0.000	sig.
	char-n-gram SVC	$+0.082_{\pm 0.008}$	0.000	sig.

Paired bootstrap with 1000 resamples, computed within each learning rate. Δmacro-F1 is arabizi − other (positive = arabizi better), $p_{\max}$ is the largest p-value across the five seeds. Verdict: "sig." if $p_{\max} < 0.05$, "borderline" if $0.05 \leq p_{\max} < 0.10$, otherwise "n.s."

5.6.3 Text length

While the performance is similar for short and medium comments (macro-F1 ≈ 0.55 up to 8 words, 0.527 at 9–20 words), it drops for comments longer than 20 words. With 0.455 macro-F1 and 0.555 accuracy, unlike the previous section, the outcome is not related to class composition. Only 63% of the comments are positive, therefore not more than average. The 128-token truncation is also unlikely to be the reason - at the average 1.44 subwords per word, a comment would need around 90 words to reach the limit. The result is probably caused by the small size of the slice (142 comments) or their mixed sentiment.

5.7 Tokenizer fit

In Section 2 I argued that a tokenizer trained on the wrong data over-fragments Arabizi. Figure 7 represents the performance of each tokenizer over the whole corpus (18,433 words). Both DarijaBERT variations split Arabizi words into fewer subwords than the other tokenizers - 1.44 subwords per word for DarijaBERT-arabizi and 1.51 for DarijaBERT-mix. All of the other tokenizers split words into more than 2 subwords on average (from 2.16 (XLM-R) through 2.32 (mBERT) to 2.63 (MARBERTv2)). The splits have no correlation with vocabulary size. DarijaBERT-arabizi splits into the least subwords on average, with only a 110k vocabulary, less than half of XLM-R's 250k. The examples in Table 4 support the outcomes. The Darija tokenizers keep words such as wa3er, 3ndi, and l9hwa whole with their digits attached, whereas the other tokenizers break them into characters and split the digits off.

It is worth highlighting that the unknown-token rate is negligible for every tokenizer (below 0.01%), except for MARBERTv2 whose rate reaches 0.27%. Therefore, information is not lost through fragmentation. Each encoder sees the whole input, but the non-DarijaBERT tokenizers split each word more. That results in smaller, less informative pieces, diluting the per-word

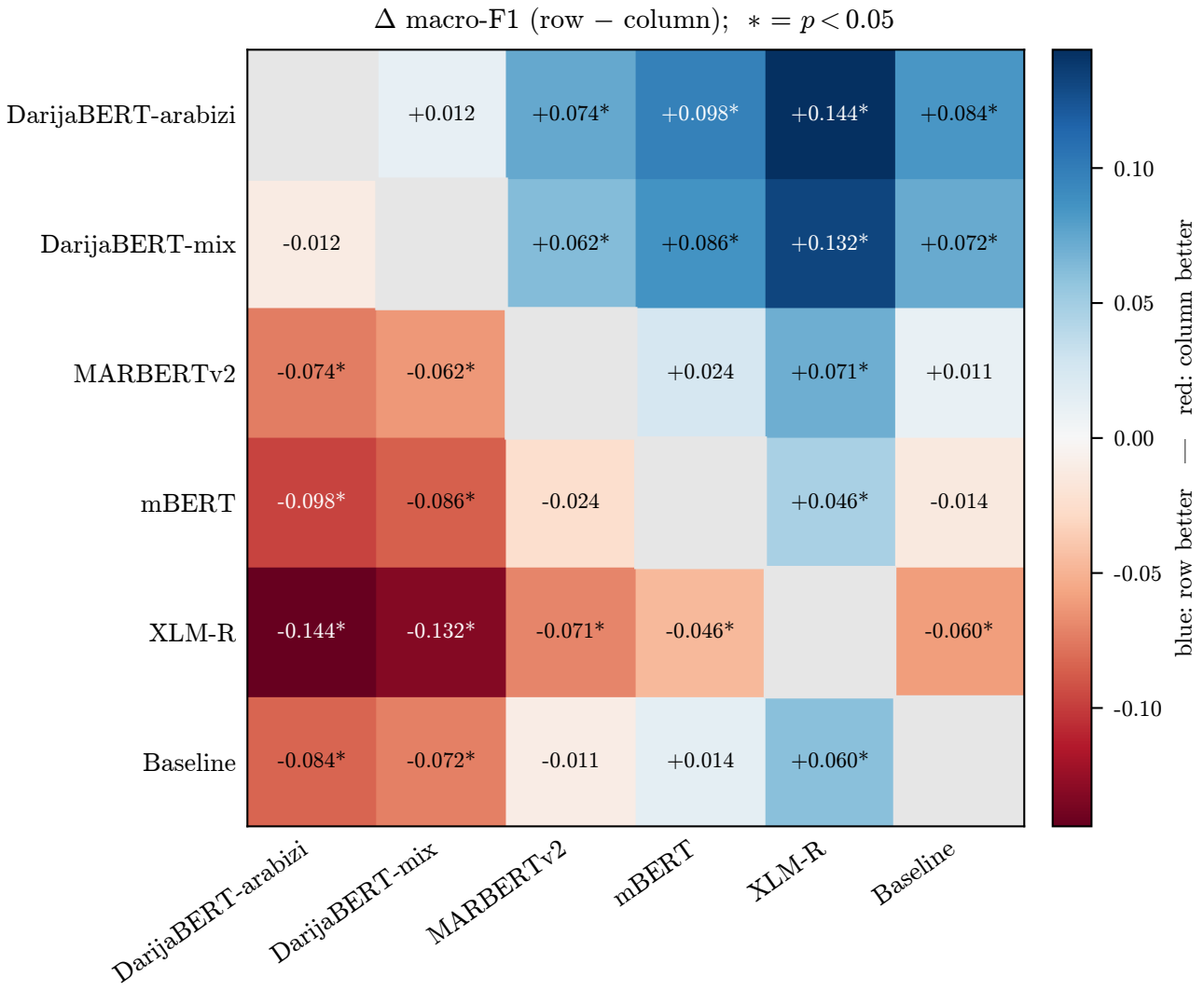

Figure 5: Pairwise differences in macro-F1 (row − column) at lr= 2×10^{-5} from a 1000-iteration paired bootstrap, $*$ marks $p < 0.05$. DarijaBERT-arabizi's advantage over every non-DarijaBERT model is significant, while it is tied with DarijaBERT-mix.

signal [34].

The tokenizer fit does not fully align with the previous outcomes. The two best-fitting tokenizers belong to the two best performing models. Therefore, fragmentation cleanly separates the Darija encoders from the rest. However, beyond the best models it does not track the overall outcomes. XLM-R has the lowest fragmentation of the non-Darija models (2.16) yet the worst macro-F1. On the other hand, MARBERTv2 has the highest fragmentation (2.63) yet outscores both XLM-R and mBERT.

6 Discussion

The model ordering shows how closely each model's pretraining matches the dialect. The two DarijaBERT variants lead. Almost all of remaining encoders - MARBERTv2, mBERT, and the char-n-gram baseline - have performed similarly to each other, but significantly worse than the DarijaBERT models. Their outcomes are relatively similar and their ranking order changes based on the metric we choose. The very worst model was the XLM-R, performing the weakest

Table 11: Error analysis for DarijaBERT-arabizi at lr= 2×10^{-5} along Darija-specific axes.

Slice	n	Macro-F1	Accuracy
no code-switch	12,630	0.552	0.640
code-switch	1,310	0.536	0.831
digits: none	7,105	0.561	0.676
digits: low	3,935	0.541	0.667
digits: high	2,900	0.514	0.605
length $\leq$3 (very short)	5,880	0.549	0.661
length 4–8 (short)	4,855	0.553	0.675
length 9–20 (medium)	2,495	0.527	0.649
length >20 (long)	710	0.455	0.555

Predictions pooled over the 5 seeds, n is the number of examples per slice. Accuracy rises sharply on code-switched text while macro-F1 does not, as that slice is overwhelmingly positive. Long texts degrade both metrics.

with a clear gap.

The Darija models perform best not because they are larger. Each encoder follows the same ≈86-million-parameter architecture, which means that they differ only in pretrained weights and vocabulary. That is shown by the 170M-parameter DarijaBERT-arabizi outperforming the biggest model included, 278M-parameter XLM-R. The performance difference comes from two separate sources, both related to specialization.

Firstly, the Darija tokenizers split words into roughly 1.4–1.5 subwords, while the rest of the tokenizers span 2.2–2.6 subwords per word. Therefore, the model receives a coherent unit, rather than a string of multiple fragments [34].

Secondly, the correlation between pretraining data and the targeted dialect is important. The tokenizer fit affects the ranking, but does not fully explain it (XLM-R has the lowest fragmentation, yet is outperformed by MARBERTv2, with the highest fragmentation). Therefore, the difference that makes DarijaBERT models perform best is their Arabizi-aware tokenizer and dialect-specific pretraining data. This indicates that, in this case, specialization matters more than scale.

The rankings based on different metrics disagree. Taking into account the accuracy, the char-n-gram baseline performs very well. At 0.649 it is the second-highest of any model, behind only DarijaBERT-arabizi (0.658). What is important, none of the models reach the always-positive floor of 0.669, which would be obtainable by blindly predicting the majority of positive class. Therefore, if we ranked classifiers only by accuracy, we would prefer a classifier that does nothing other than that. The illusion is exposed by chance-corrected agreement. The baseline's κ of 0.228 is significantly below DarijaBERT-arabizi's 0.366, with a roughly 60% higher agreement. Its minority-class F1 scores are among the lowest outcomes. Figures 8 and 9 visualize the duality of baseline's high accuracy and low chance-corrected metrics. This shows that it earns accuracy almost entirely from the majority class.

The Darija-specific error axes separate genuine model properties from properties of the data itself. The phonemic digit density axis, which we analyzed shows that although performance of all explored models declines while phonemic digit density rises, the DarijaBERT models perform the best at every level. The Arabizi-aware tokenizer helps in that case, as it keeps the subwords intact, instead of separating the digits. The score of around 0.51 macro-F1 at the highest density, against at most 0.44 for the others illustrates this effect.

The two remaining axes tell us more about the dataset than about models. The 86% positive

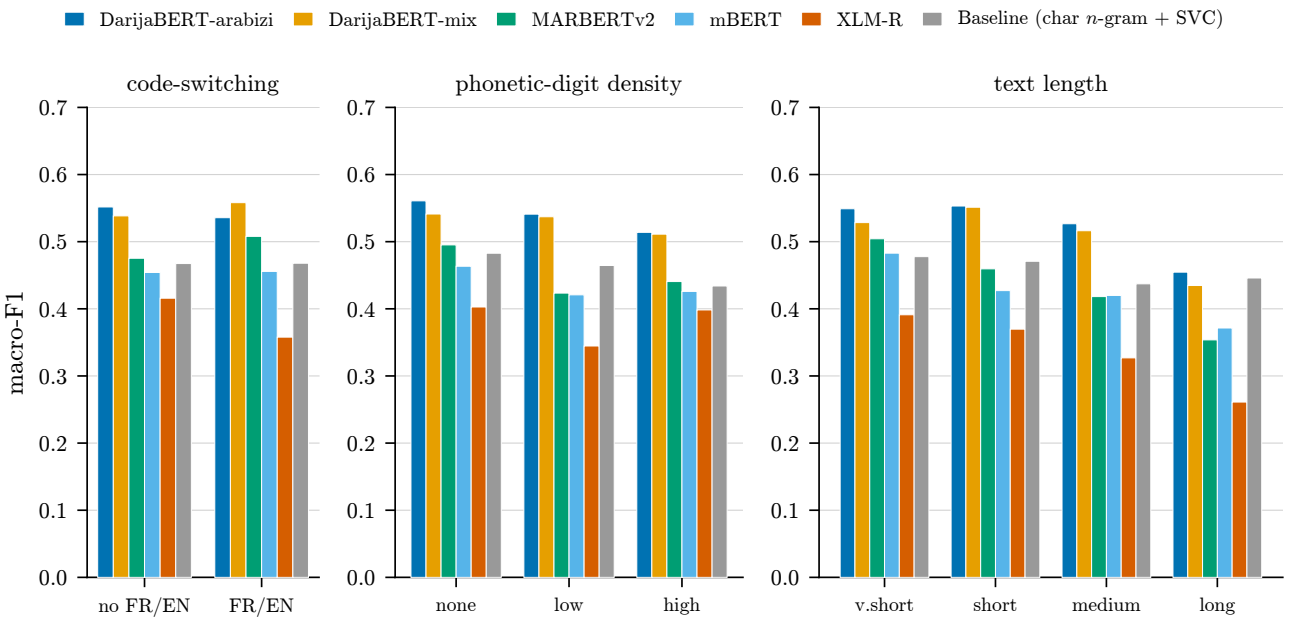

Figure 6: Macro-F1 broken down by Darija-specific axes at lr= 2×10^{-5}: code-switching presence, phonemic-digit density, and text length.

code-switched subset caused a spike in accuracy, but brought no gain in macro-F1. The bad performance on the axis of comment length affected all models alike, suggesting the impact of mixed-sentiment and small size of that slice, rather than specific model-related issues.

Lastly, the case of XLM-R is worth mentioning. Being both the largest model and the worst performer, its failure is an interesting observation. At the highest learning rate multiple of its runs collapsed onto single classes. This specific behavior was already observed when large pretrained models were fine-tuned on small imbalanced datasets [29, 15, 36]. Despite having the most parameters and the lowest fragmentation of all non-Darija models XLM-R was the least stable and weakest for the task.

The limitations of experiments conducted for this thesis qualify these findings. The heavily imbalanced, relatively small corpus annotated by only two annotators was the main limitation. The fixed protocol approach also enforced several limitations. To make the outcomes comparable, the batch size and learning rates were the same for all models. Adjudication was conducted using random choice instead of a third native speaker, therefore a portion of gold labels is uncertain. The moderate inter-annotator agreement also qualifies the absolute scores. With Cohen's κ of 0.219 between annotators, the gold labels themselves carry disagreement, and agreement at this level tends to impose a ceiling on measurable model performance: a classifier is scored against labels that two native speakers would not reproduce identically. This ceiling is expected to bind hardest where the annotators disagreed most - the negative–neutral boundary - and that is indeed where the models make most of their errors (Section 5.4). Part of the residual error of even the best model is therefore attributable to genuine ambiguity in the labels rather than to the encoders themselves. The study also covers only a single domain of comments, with no cross-domain evaluation. Finally, the task itself abstracts away context that carries sentiment. The comments were written in reaction to videos that neither the annotators nor the models see, and preprocessing removed emojis - the written counterpart of non-verbal cues. The same sentence can express opposite sentiments depending on such cues, so reducing inherently multimodal communication to unimodal text discards information relevant to sentiment. This does not affect the comparison between models, since all of them receive the same text, but it is

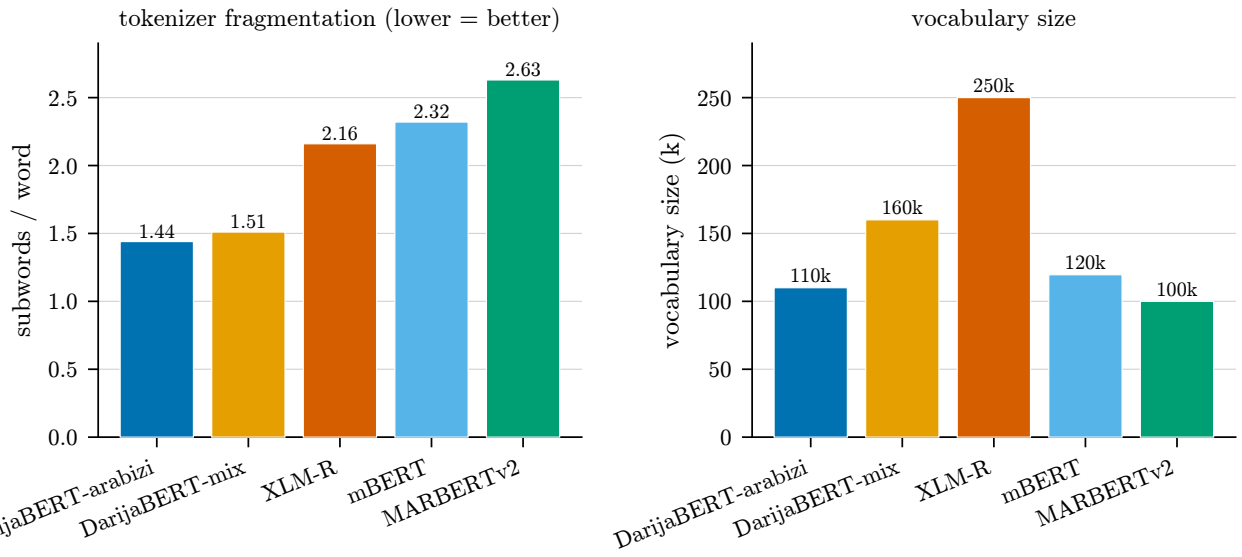

Figure 7: Tokenizer fit on Darija Arabizi, measured over the whole corpus. The Arabizi-specific DarijaBERT tokenizers produce far fewer subwords (≈1.4–1.5 subwords/word) than the multilingual and MSA-oriented tokenizers (≈2.2–2.6) despite smaller vocabularies.

a general limitation of text-only sentiment analysis that qualifies the absolute scores.

7 Conclusions and Further Research

The results provide clear answers to RQ1 and RQ2. DarijaBERT-arabizi achieved the highest macro-F1 scores consistently at every learning rate. It performed the best across the five seeds and five folds, having a per-rate spread of only 0.007. The advantage it has over every non-DarijaBERT model is significant, except for the second DarijaBERT variation used in this thesis, the DarijaBERT-mix. They are statistically tied, therefore both of these models can be qualified as the best performers. In the context of RQ1 that ranking would hold even for default settings, as the DarijaBERT-arabizi was proven to be practically insensitive to learning rate.

The answer to RQ3 can be derived clearly from the data. Not all transformers beat the baseline - the generic multilingual and Arabic encoders perform equally or even worse dependent on the metric. However, the DarijaBERT models outperform it clearly, but the visibility of that difference is dependent entirely on the metric. That is why macro-F1 and kappa are reported in addition to accuracy.

To answer RQ4 it is important to notice that dialect-specific pretraining shows up in two places. In tokenization, the Arabizi-aware tokenizers split words into far fewer subwords on average than the others. In the error patterns, where the Darija models keep their advantage as phonemic-digit density rises, confusing mostly adjacent sentiment classes, rather than the polar opposites.

In conclusion, the Darija-specific models performed the best on sentiment analysis of Moroccan Darija Arabizi. DarijaBERT-arabizi was the strongest model, statistically tied with the mixed variation. These models were significantly ahead of the larger multilingual and Arabic encoders, making the main conclusion state that for this task, specialization outperforms scale. The DarijaBERT-arabizi is therefore recommended for this task, even with default hyperparameters.

The limitations mentioned above point towards obvious next steps. The main focus in

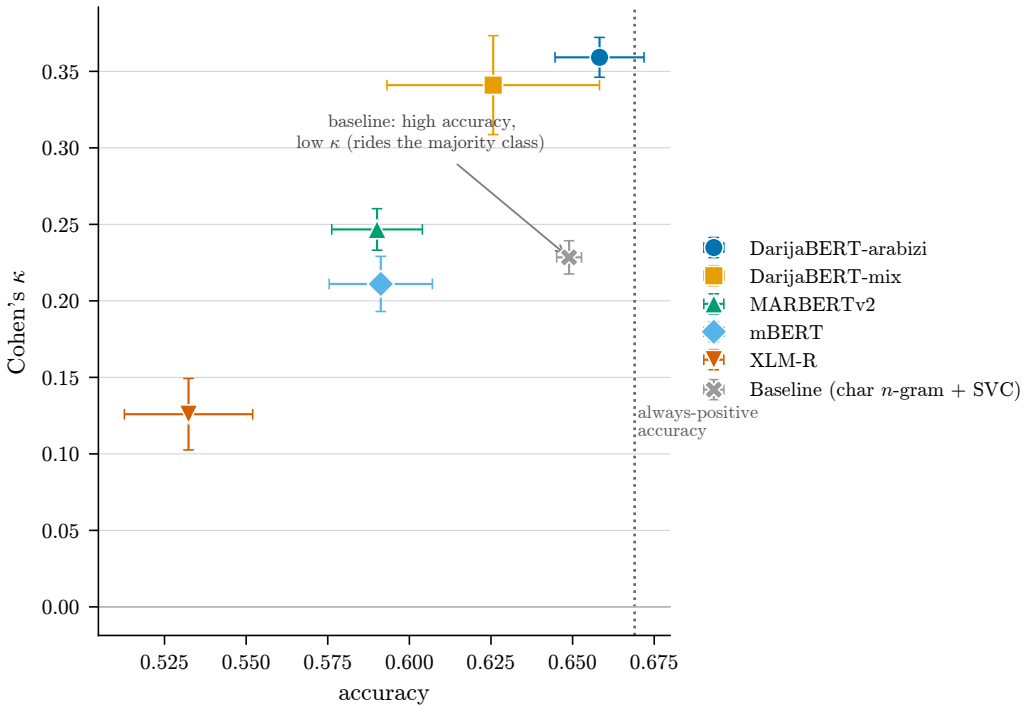

Figure 8: The imbalance trap at lr= 2×10^{-5}. The char-n-gram baseline matches the transformers on accuracy but scores far lower on chance-corrected κ; like every model, its accuracy falls just short of the always-positive line (dashed).

further work should be a larger, more balanced dataset, with stronger adjudication, which would reduce variance. In the context of the models, learning rate tuning and warmup per model could improve performance, specifically of XLM-R. Further work should also include domain-adaptive pretraining of DarijaBERT-arabizi on additional sources. A comparison against prompted multilingual large language models would help assess how well dialect-specialized encoders can perform.

Acknowledgements

I am deeply grateful to my supervisors, Suzan Verberne and Max van Duijn, for their guidance, patience, and thoughtful feedback throughout this thesis. I would also like to thank Lilia Moussafi and Hiba Benlkaid for their careful re-annotation of the Moroccan Darija dataset. Their linguistic expertise and dedication were essential to this work.

References

[1] Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. DziriBERT: a pre-trained language model for the Algerian dialect. *arXiv preprint arXiv:2109.12346*, 2021.

[2] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of ACL-IJCNLP 2021 (Volume 1: Long Papers)*, pages 7088–7105. Association for Computational Linguistics, 2021.

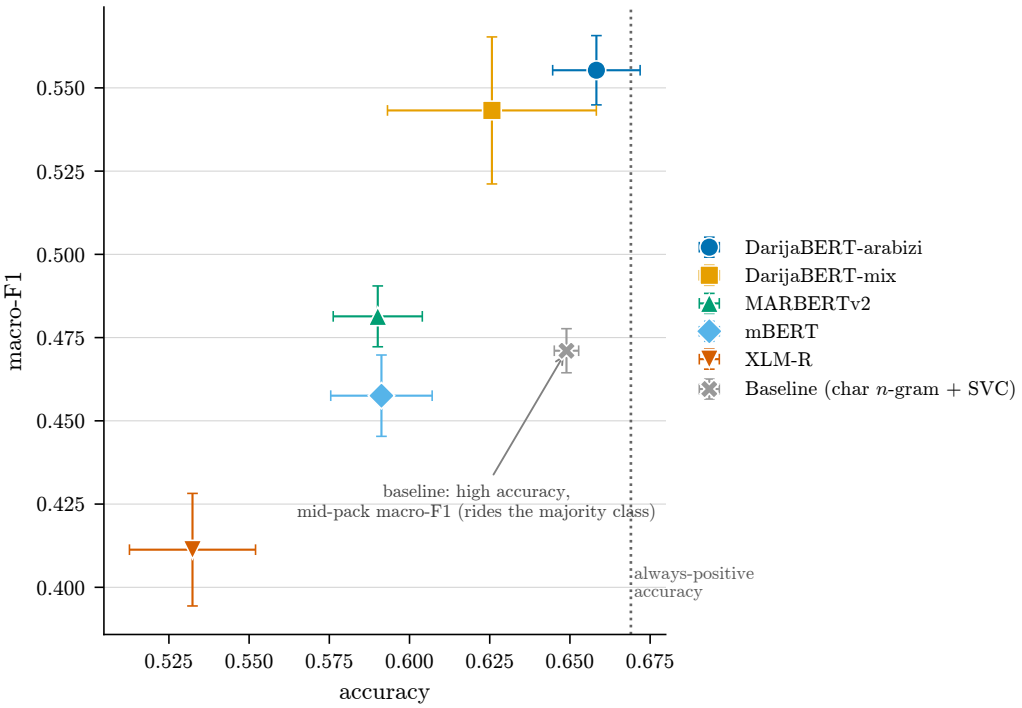

Figure 9: Accuracy versus macro-F1 at lr$= 2 \times 10^{-5}$ (error bars are ± 1 s.d. across seeds). The char-n-gram baseline lies far to the right on accuracy - close to the always-positive line - yet only mid-pack on macro-F1, level with MARBERTv2 and mBERT despite their much lower accuracy, while the dialect-specific models lead on both axes.

[3] Amani A. Aladeemy, Ali Alzahrani, Mohammad H. Algarni, Saleh Nagi Alsubari, Theyazn H. H. Aldhyani, Sachin N. Deshmukh, Osamah Ibrahim Khalaf, Wing-Keung Wong, and Sameer Aqburi. Advancements and challenges in arabic sentiment analysis: A decade of methodologies, applications, and resource development. *Heliyon*, 10:e39786, 2024. Comprehensive survey of Arabic sentiment analysis, including deep learning and transformer models.

[4] Meriem Amnay, Abdelfattah Toulaoui, Mourad Jabrane, and Imad Hafidi. Unveiling the effectiveness of contextual embeddings in sentiment analysis on moroccan darija: A comparative study with traditional techniques. *IAENG International Journal of Computer Science*, 52(10):1–10, 2025. Comparative study of transformer models vs. traditional methods for Moroccan Darija sentiment.

[5] Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT)*, pages 9–15. European Language Resources Association, 2020.

[6] Gilbert Badaro, Ramy Baly, Hazem Hajj, Wassim El-Hajj, Khaled Bashir Shaban, Nizar Habash, Ahmad Al-Sallab, and Ali Hamdi. A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3):1–52, 2019.

[7] Nawfal Benhamdane. Sentiment analysis for darija (arabic dialect). Hugging Face model, 2025. Community BERT model for binary sentiment classification in Moroccan Darija.

[8] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In *Proceedings of EMNLP-CoNLL 2012*, pages 995–1005. Association for Computational Linguistics, 2012.

[9] ElMehdi Boujou, Hamza Chataoui, Abdellah El Mekki, Saad Benjelloun, Ikram Chairi, and Ismail Berrada. An open access NLP dataset for Arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000*, 2021.

[10] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[11] Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968.

[12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics, 2020.

[13] Kareem Darwish. Arabizi detection and conversion to Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 217–224. Association for Computational Linguistics, 2014.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[15] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

[16] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. Association for Computational Linguistics, 2018.

[17] Sara El Ouahabi, Safâa El Ouahabi, and El Wardani Dadi. Multi-domain dataset for Moroccan Arabic dialect sentiment analysis in social networks. In *2023 IEEE 6th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, pages 01–07, 2023.

[18] Mouaad Errami, Mohamed Amine Ouassil, Rabia Rachidi, Bouchaib Cherradi, Soufiane Hamida, and Abdelhadi Raihani. Sentiment analysis on moroccan dialect based on machine learning and social media content detection. *International Journal of Advanced Computer Science and Applications*, 14(3), 2023.

[19] Kamel Gaanoun, Naira Abdou Mohamed, Anass Allak, and Imade Benelallam. DarijaBERT: a step forward in NLP for the written Moroccan dialect. *International Journal of Data Science and Analytics*, 2024.

[20] Moncef Garouani and Jamal Kharroubi. MAC: An open and free Moroccan Arabic corpus for sentiment analysis. In *Innovations in Smart Cities Applications Volume 5 (Proc. 6th Int. Conf. on Smart City Applications)*, volume 393 of *Lecture Notes in Networks and Systems*, pages 849–858. Springer, Cham, 2022.

[21] Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP)*, pages 92–104. Association for Computational Linguistics, 2021.

[22] Mouad Jbel, Imad Hafidi, and Abdulmutallib Metrane. MYC: A Moroccan corpus for sentiment analysis. In *Advances in Machine Intelligence and Computer Science Applications (ICMICSA 2022)*, volume 656 of *Lecture Notes in Networks and Systems*, pages 59–68. Springer, Cham, 2023.

[23] Mouad Jbel, Mourad Jabrane, Imad Hafidi, and Abdulmutallib Metrane. Sentiment analysis dataset in Moroccan dialect: bridging the gap between Arabic and Latin scripted dialect. *Language Resources and Evaluation*, 59:1401–1430, 2025.

[24] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395. Association for Computational Linguistics, 2004.

[25] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP 2018: System Demonstrations*, pages 66–71. Association for Computational Linguistics, 2018.

[26] Yassir Matrane, Faouzia Benabbou, and Nawal Sael. A systematic literature review of arabic dialect sentiment analysis. *Journal of King Saud University – Computer and Information Sciences*, 35(6):101570, 2023. Systematic review of resources, methods, and challenges in Arabic dialect sentiment analysis.

[27] Abir Messaoudi, Hatem Haddad, Chayma Fourati, Moez Ben Haj Hmida, Aymen Ben Mabrouk, and Mohamed Graiet. TunBERT: Pretrained contextualized text representation for Tunisian dialect. *arXiv preprint arXiv:2111.13138*, 2021.

[28] Soukaina Mihi, Brahim Ait Ben Ali, Ismail El Bazi, Sara Arezki, and Nabil Laachfoubi. MSTD: Moroccan sentiment Twitter dataset. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(10):363–372, 2020.

[29] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations (ICLR)*, 2021.

[30] Monsif Nadir. Darijabert for sentiment analysis. Hugging Face model, 2025. Community DarijaBERT-based sentiment classifier (three-way).

[31] Abdusalam F. Ahmad Nwesri, Nabila Almabrouk S. Shinbir, and Amani Bahlul Sharif. Sentiment analysis on Arabic dialects: A multi-dialect benchmark. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHASIS, RANLP 2025)*, pages 86–91, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, Bulgaria.

[32] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.

[33] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[34] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of ACL-IJCNLP 2021 (Volume 1: Long Papers)*, pages 3118–3135. Association for Computational Linguistics, 2021.

[35] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics, 2016.

[36] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. Revisiting few-sample BERT fine-tuning. In *9th International Conference on Learning Representations (ICLR)*, 2021.

A Annotation guidelines

Neutral (0). Mostly informative, objective phrases - making a statement, stating a fact, or providing scientific material. The language is devoid of emotional features such as subjective adjectives and "I feel" statements; however, some "I feel" statements can still be neutral (e.g. "I feel cold"). *Examples:* "One plus one equals two."; "The earth is round, but not a perfect sphere."

Non-neutral. A way to express personal viewpoints, opinions, experiences, and biases. As a rule it always carries a touch of one's own preference, which is either:

- **Positive (+1):** expressions of endearment, or subjective opinions stating satisfaction, amazement, affection, or excitement. *Example:* "I love the beach!"
- **Negative (−1):** dissatisfaction, disappointment, disgust, frustration, regret, insecurity, or shame. *Example:* "I hate cold weather!"

Nuance. Context is valuable, especially for slang, sarcasm, and second-degree comments. For example, "ugh! what a short movie, it was only 3 hours long." conveys frustration with the length even though no negative language is used, and would be labelled −1.