Improving on State-of-the-Art Models for Sentiment Analysis on Saudi-English Code-Switching Text

Samaher Alghamdi[1,2], Paul Rayson[2], and Reem Alotaibi[1]

[1]Department of Information Technology, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia
[2]School of Computing and Communications, Lancaster University, Lancaster, UK

Abstract

Inserting English words, phrases, or sentences while writing or speaking in the Saudi Arabic dialect has become a widespread phenomenon in Saudi society. This phenomenon is linguistically called code-switching. It remains unclear how current sentiment analysis methods perform on Saudi-English code-switching text. In this paper, we address this gap by conducting the first sentiment analysis study on Saudi-English code-switching text. We present the first Saudi-English Sentiment Analysis Code-Switching Dataset (SESA-CSD) and establish baseline results on this dataset. By evaluating multiple state-of-the-art small language models, we achieve improvements over the baseline of 3% to 11% in both accuracy and macro-F1. Among all small language models, XLM-RoBERTa achieved the highest performance, with an accuracy of 95.50% and a macro-F1 of 95.53%. Our findings indicate that multilingual and Arabic small language models, such as XLM-RoBERTa, GigaBERT, and SaudiBERT, consistently outperform bilingual Arabic-English large language models, such as Fanar and ALLaM, across zero-shot and multiple few-shot settings.

1 Introduction

Sentiment analysis is the computational analysis of people's opinions, attitudes, and behaviors towards a given topic, issue, or entity (Liu, 2012). It is a well-established task within Natural Language Processing (NLP) and has been extensively studied for decades across diverse languages, including English and Arabic. The Arabic language poses considerable challenges for many NLP tasks due to its rich morphological system, substantial dialectal variation, orthographic inconsistencies, and limited availability of linguistic resources (Darwish et al., 2021; Badaro et al., 2019). Nevertheless, significant research efforts have been undertaken to address these challenges, and continuous progress has been made in this area (Badaro et al., 2019).

The Saudi dialect is one of the Arabic dialects spoken in Saudi Arabia. Several studies have been conducted on sentiment analysis in Saudi dialects and have reported promising results (Addawood et al., 2020; Al-Rubaiee et al., 2016; Almuqren and Cristea, 2021; Bayazed et al., 2020; Alqahtani et al., 2022). However, a noticeable shift in the language used in everyday communication among speakers of the Saudi dialect has become evident (Alowidha, 2024). Recently, it has become common practice to insert an English word, phrase, or sentence when speaking or writing in the Saudi dialect, a phenomenon linguistically known as code-switching. According to Poplack (1980), code-switching is the shifting between two languages, which can occur within the same sentence or between sentences. It is categorized into three main types: intra-sentential, when switching occurs within a sentence; inter-sentential, when switching occurs between sentences; and tag-switching, when tags such as interjections, fillers, or idioms (Ternovykh and Nikiforova, 2023) are inserted into a sentence. Intra-word code-switching is another form presented by Stefanich et al. (2019), which occurs at the word level by adding a root or affix from one language to another.

Code-switching has become increasingly prevalent among bilingual speakers of English and Arabic in Saudi Arabia, particularly among younger generations. Some examples of code-switching in the Saudi dialect are presented in Table 1. With Saudi Vision 2030, which encourages education and scholarships (Vision2030, 2025), the number of Saudis who speak English is expected to increase, leading to more code-switching in society. With 99% of Saudis using the Internet (Communications,Space and Technology Commission, 2025), we also expect the volume of generated and

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 218–228

Code-switching Type	Example	Translation
Intra-sentential	اليوم عندنا meeting مع المدير الجديد.	Today we have a meeting with the new manager.
Inter-sentential	I will get a break.أنا اليوم مرة تعبان	I am so tired today. I will get a break.
Tag switching	Oh my god, مرة حلو الفستان	Oh my god, the dress is gorgeous.
Intra-word	اليوم بنروح ال mall .	Today we are going to the mall.

Table 1: Examples of Saudi-English code-switching texts.

shared data to be increasingly influenced by this phenomenon.

While the sentiment analysis of the Saudi dialect has been studied, no prior study has examined sentiment analysis of Saudi-English code-switching texts. Therefore, it is unclear how existing methods and models perform on this type of textual data, highlighting the importance of investigating this phenomenon computationally. Our contribution can be summarized as:

1. We conducted the first sentiment analysis study on Saudi–English code-switching texts and established baseline results to support future research.

2. We present the first Saudi–English code-switching dataset for sentiment analysis called Saudi-English Sentiment Analysis Code-Switching Dataset (SESA-CSD), obtained by annotating an existing dataset, which will be publicly available.

3. We show improvements over the baseline by testing multiple state-of-the-art multilingual and Arabic language models.

The rest of the paper is organized as follows: Section 2 discusses related work on code-switching sentiment analysis. Section 3 provides details on the dataset, including sampling, annotations, and cleaning. Section 4 presents the experimental details and results, and Section 5 provides findings and conclusions.

2 Related Work

With respect to sentiment analysis of Saudi-English code-switching texts, no prior work has examined sentiment analysis in this type of textual data. In contrast, the literature includes multiple studies on sentiment analysis of code-switching text in other Arabic dialects and other languages. In this section, we present some related work on sentiment analysis of code-switching text in which English serves as the secondary language for switching.

2.1 Sentiment Analysis of Arabic Code-Switching Text

Most studies on sentiment analysis of code-switching in Arabic have focused on North African Arabic dialects, including Moroccan Arabic, Tunisian Arabic, Algerian Arabic, and Egyptian Arabic. This can be explained by the nature of these dialects, which are known for extensive code-switching.

Adouane et al. (2020) showed that in Algerian YouTube comments, a CNN architecture outperforms LSTM, BiLSTM, and SVM models, achieving a 60.17% macro-F1. Adding sentiment lexicons to the CNN and augmenting data for minority classes improved the CNN's performance on those classes. Similarly, Almasah et al. (2023) proposed two CNN architectures to analyze the sentiment of 200 Egyptian-English reviews. The proposed CNN architectures outperform LSTM, BiLSTM, and a hybrid architecture that combines the aforementioned models, achieving an accuracy of 83%. Furthermore, in Tunisian Facebook comments, Jerbi et al. (2019) applied LSTM, BiLSTM, Stacked LSTM, and Stacked BiLSTM, and showed that both Stacked LSTM and Stacked BiLSTM surpassed the other models, achieving accuracies of 90% and 88%, respectively.

Regarding transformer-based models, Boudad et al. (2023) investigated the performance of multilingual pre-trained language models such as multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020), and Arabic pre-trained language models such as AraBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2021), QARIB (Abdelali et al., 2021), CAMeL (Inoue et al., 2021), and DarijaBERT (Gaanoun et al., 2025) in sentiment analysis on Moroccan datasets and an Arabic multi-dialect code-switching dataset that in-

cludes a mixture of English and French languages. In code-switching data, the transformers outperformed traditional machine learning and deep learning models, while all models achieved comparable performance, with F1 score ranging from 81.32% to 82.44%.

A recent study by Sherif and Sabty (2024) analyzed sentiment in a dataset of 4,100 Egyptian–English YouTube comments. They employed a BiLSTM with a self-attention layer and a hybrid transformer model and applied multiple Arabic and Arabic–English embeddings. By applying an ensemble approach, feeding the outputs of the two best models into a single hidden layer, they achieved an F1 score of 92.54%. Moreover, they compared the performance of several large language models, such as GPT-3.5-turbo, Gemini-1.0-pro, Gemini-1.5-pro, and GPT-4o, across multiple configurations. Fine-tuning GPT-3.5-turbo outperformed all models and achieved a comparable result to the ensemble model, with an F1 score of 92.76%.

2.2 Sentiment Analysis of Non-Arabic Code-Switching Text

Spanish–English and Hindi–English are the most common language pairs that have been studied in the sentiment analysis literature. SemEval-2020 shared task 9 covered the sentiment analysis of code-mixed tweets in these two language pairs (Patwa et al., 2020). Multiple proposed models were based on the transformer architecture. The best performing model in the Hindi–English pair was the XLM model (Conneau and Lample, 2019) trained using adversarial samples to improve regularization, achieving a 75% F1 score (Liu et al., 2020). In the Spanish–English language pair, the best model achieved an 80.6% F1 score by augmenting machine translation data and using XLM embeddings as input to a fully connected layer, while optimizing the weighted loss based on the complexity of the code-mixed data (Ma et al., 2020).

In Tamil-English and Malayalam-English code-switching texts, Balouchzahi and Shashirekha (2021) showed that an ensemble model combining Multilayer Perceptron, eXtreme Gradient Boosting, and Logistic Regression trained on Char sequences, Byte Pair Encoding subwords, and syntactic n-gram features outperforms both the sequential neural network model and the transfer learning model, achieving an F1 score of 72% and 62% in Tamil-English and Malayalam-English, respectively. In Hindi-English code-switching texts, Lal et al. (2019) utilized two encoders built on a BiLSTM architecture to capture the sentiment of both the whole sentence and individual subwords sentiments. The encoders' outputs are combined with additional linguistic features and passed to fully connected layers for the final prediction, achieving an F1 score of 82.7%.

In the context of transformer-based models, Sharma et al. (2023) employed the logits from two transformer models—BERT, mBERT, and XLM-R—and fed them into a fully connected neural network for final classification. The proposed approach was applied to the English–Hindi and English–Spanish language pairs and demonstrated substantially higher performance compared with the reported results on the GLUECoS benchmark.

3 Data

To conduct our experiments, we used the Ar-En Code-Switching Textual Dataset (ArE-CSTD) (Alharbi et al., 2024). It is one of the limited datasets available for Saudi-English code-switching text. The data was derived by the National Centre for Artificial Intelligence at the Saudi Data and Artificial Intelligence Authority (SDAIA). It is synthetic data generated by GPT-4 and contains three different versions of code-switching texts: Modern Standard Arabic with English, Egyptian dialect with English, and Saudi dialect with English. We constructed a 1k-sample dataset from the Saudi-English version and annotated it with three sentiment labels: positive, negative, and neutral. The newly constructed dataset, referred to as **Saudi-English Sentiment Analysis Code-Switching Dataset (SESA-CSD)**, constitutes the first resource for sentiment analysis of Saudi-English code-switching texts. The dataset is publicly available for research purposes [1].

3.1 Data Sampling and Annotation

The original dataset comprises 100k samples in the training set and 10k samples in the test set. To construct the SESA-CSD train set, 800 samples were selected from the training dataset based on the ratio of English words to Arabic words. All high ratios were excluded, as the dominant language was intended to be the Saudi dialect. Stratified sampling

[1] https://github.com/samaherSG/SESA-CSD

No. of sentences	1K
No. of tokens	10896
No. of Arabic words	5730
No. of English words	3878
Avg. English words per sentence	3.89
Avg. word per sentence	10.9
No. of unique words	785
Avg. ratio of English to Arabic words per sentence	0.86
Code-Mixing Index (CMI) (Gambäck and Das, 2014)	30.29
Multilingual Index (MI) (Guzmán et al., 2017)	0.25

Table 2: SESA-CSD Statistics.

Sentence	Sentiment
الحفلة كانت amazing والأجواء كانت مرّة حماس! The party was amazing, and the vibe was really exciting!	Positive
we're going to the mall ، اليوم بعد المدرسة Today after school, we're going to the mall.	Neutral
اليوم seriously كان تعب ! Today, seriously, was tiring!	Negative

Table 3: Examples of SESA-CSD.

was employed to preserve the original data distribution. As for the test set, 200 samples were obtained based on the training distribution. It is worth noting that, upon examination of the samples, some included Modern Standard Arabic (MSA) and other dialects rather than Saudi. Since SESA-CSD represents Saudi-English code-switching, those samples were replaced while maintaining the distribution. Table 2 summarizes the SESA-CSD statistics.

Regarding the SESA-CSD annotation, three Saudi native speakers and fluent English speakers annotated the dataset using three labels, yielding 268 positive, 276 negative, and 456 neutral samples. Some examples are presented in Table 3.

Inter-annotator agreement was assessed using Fleiss' kappa (Fleiss, 1971) and Krippendorff's alpha (Krippendorff, 2011), both yielding 0.82, demonstrating a high level of agreement among annotators.

3.2 Data Preprocessing

A significant step in constructing a robust model is to clean and prepare the data. Therefore, the data underwent several preprocessing steps, including:

1. Adding spaces between words (Arabic and English), numbers, and punctuation for improved tokenization. Additionally, we observed that spaces between words and the Arabic letter و – referred to as waw – were often incorrect. This Arabic letter can function either as part of a word or as a conjunction; therefore, we identified all words beginning with و and inserted spaces where necessary.
2. Normalizing Arabic and English letters.
3. Removing diacritics, numbers, invisible Unicode characters, and punctuation except for ! and ?.
4. Converting uppercase English letters to lowercase, unless they represent proper nouns.
5. Expanding English contractions.
6. Correcting certain misspelled words in Arabic.

All the aforementioned preprocessing steps were applied to the data used in training machine learning models. However, for fine-tuning small language models, fewer steps were applied to preserve sentence semantics and syntactic structure, including adding spaces, removing invisible Unicode characters, and correcting some Arabic words.

4 Experiments

In order to evaluate how current methods for sentiment analysis perform in Saudi-English code-switching text, we conducted three experiments using classical Machine Learning models, small language models (SLMs), and large language models (LLMs). To avoid confusion, we use the term SLMs for models with parameter sizes in the millions and LLMs for models with parameter sizes in the billions. The machine learning experiment is considered our baseline, and we subsequently improve upon it. In the following sections, we present the experiments in detail.

4.1 Machine Learning Models

In this experiment, we trained eight machine learning models used in the literature for sentiment analysis of Saudi dialects. The models are Support Vector Machine (SVM) with linear and nonlinear kernels, Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Gaussian Naïve Bayes (GNB), Multinomial Naïve Bayes (MNB), and K-Nearest Neighbour (KNN). All models were

trained using the scikit-learn library with default parameters and a 5-fold cross-validation. As for feature extraction, we extracted n-grams with n=1, 2, and 3, and Term Frequency-Inverse Document Frequency (TF-IDF) with n=1, 2, and 3. We repeated the experiment twice: with the code-switching sentences and after removing English words. We reported the accuracy and macro-F1 score for all experiments on the test dataset. Table 4 presents the results for the experiment in the code-switching sentences.

By analysing the results in Table 4, it is evident that the highest performance was achieved by a linear SVM with unigram features, attaining an accuracy of 84% and a macro-F1 of 82.83%. In general, the linear SVM exhibits strong performance across multiple feature sets, although some models achieve comparable performance for certain features. Notably, all models perform better with unigram features and with combined n-grams, while performance degrades when using bi-gram and tri-gram features. To further improve the results, we applied a voting classifier that combines LR, a linear SVM, and GNB on TF-IDF with n=1, 2, and 3. Both accuracy and macro-F1 improved by 1% and 1.77%, respectively, reaching 85% and 84.60%, and these results serve as our baseline.

A common practice in Arabic sentiment analysis studies is to remove all English words during the data cleaning phase to prepare the data for the models. We argue that these English words carry sentiment and convey important information that may contribute to the overall sentiment of the sentence. To test our assumption that English words carry sentiment and that their removal may affect model performance, we repeated the previous experiment on the dataset after removing all English words and reported the results in a Table 5.

Consistent with our assumption, most models exhibit a sharp decline in performance, especially those that performed strongly on the code-switching sentences. The best results in this experiment were 75% and 72.72% for accuracy and macro-F1, respectively, indicating decreases of about 9% in accuracy and 10% in macro-F1 compared with the earlier experiment. Although MNB achieved the highest performance in this experiment, LR and linear SVM performed comparably. Consistent with our previous findings, linear SVM performs strongly across multiple features, and all models perform better with unigram and with combined n-grams. A notable finding is that RF, DT, and KNN improved with some features in both accuracy and macro-F1 after removing the English words. In contrast, these models had previously demonstrated lower performance on code-switching sentences. This pattern may be attributed to the introduction of noise by English words, which these models appear particularly sensitive to.

4.2 Small Language Models

To improve upon the baseline, we fine-tuned several current state-of-the-art SLMs. Among multilingual SLMs, we evaluated mBERT and XLM-R (base and large). For Arabic SLMs, we evaluated SaudiBERT (Qarah, 2024), AraBERT (base and large), and CAMeLBERT (dialectal and mixed). In addition, we evaluated GigaBERT (Lan et al., 2020) as a code-switching model. For each model, multiple configurations with early stopping were applied to determine the optimal performance. Table 6 presents the results. It is important to note that for some models, the same results were repeated with different configurations but were not reported due to space limitations.

It is notable that all SLMs outperform our baseline, with the CAMeLBERT models as the exception. The overall improvement ranges from 3% to 11% in both accuracy and macro-F1. XLM-R models exhibit strong performance among all models, with the base model outperforming the large variant, demonstrating a strong ability to classify sentiments presented in Saudi–English code-switching, reaching 95.50% and 95.53% in accuracy and macro-F1, respectively. GigaBERT demonstrated strong performance and provides evidence that training SLMs on code-switching data is essential for NLP tasks involving code-switching text. For Arabic SLMs, all models achieved comparable performance, except for the CAMeLBERT models. Despite SaudiBERT being smaller in parameter count (143M), it achieved performance comparable to AraBERT-Large (371M). This could be attributed to it being the only model explicitly trained on the Saudi dialect, with English words retained in the training data if their proportion did not exceed 50% of the total number of words in the sentence.

Model	TF-IDF (n=1)		TF-IDF (n=2)		TF-IDF (n=3)		TF-IDF (n=1,2,3)		n-gram (n=1)		n-gram (n=2)		n-gram (n=3)		n-gram (n=1,2,3)	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LR	80.50	79.63	75.00	71.48	68.00	60.21	80.00	78.70	81.50	80.86	74.00	71.31	70.50	65.75	**83.00**	**81.83**
SVM (linear)	**82.00**	**81.07**	**76.00**	**74.04**	**70.00**	**66.13**	81.00	79.91	**84.00**	**82.83**	74.50	72.03	70.50	66.26	82.50	81.25
SVM (non-linear)	82.00	80.90	75.00	70.71	59.00	45.43	79.00	77.47	79.50	77.76	67.50	63.65	51.00	43.01	79.50	77.56
RF	74.50	71.62	69.50	63.32	55.50	38.52	76.00	73.25	76.00	73.47	68.50	62.03	56.00	37.85	73.50	70.07
DT	72.00	69.24	65.50	59.74	57.00	45.07	66.00	63.25	75.50	74.02	67.00	61.07	63.00	53.38	74.00	72.01
MNB	76.00	74.11	72.00	67.34	67.50	59.01	74.50	71.41	78.50	77.43	**76.00**	**74.18**	**71.50**	**67.92**	81.50	80.56
GNB	74.00	72.93	70.50	69.60	59.00	57.72	**82.00**	**81.43**	74.00	72.73	70.00	69.02	58.50	57.37	81.50	80.67
KNN	79.50	78.00	72.50	69.88	59.50	56.25	76.00	74.47	66.00	63.23	50.00	48.30	34.50	29.12	50.00	47.47

Table 4: Accuracy (Acc) and macro-F1 (F1) reported on in the code-switching sentences. Numbers in bold indicate the best result for each feature.

Model	TF-IDF (n=1)		TF-IDF (n=2)		TF-IDF (n=3)		TF-IDF (n=1,2,3)		n-gram (n=1)		n-gram (n=2)		n-gram (n=3)		n-gram (n=1,2,3)	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LR	73.50	**70.99**	68.00	63.10	60.50	49.73	73.00	69.94	**75.00**	**71.48**	68.00	63.57	60.50	49.92	74.50	71.40
SVM (linear)	73.00	69.66	**72.00**	**67.97**	**64.50**	**57.04**	**75.00**	**72.39**	70.00	66.36	68.50	64.21	62.50	53.14	72.00	68.25
SVM (non-linear)	**74.00**	70.32	67.50	61.55	58.00	44.97	72.50	69.09	69.50	64.97	67.00	58.72	57.00	41.06	72.50	68.76
RF	72.00	68.79	64.00	56.78	59.00	47.63	68.50	63.81	72.00	68.70	65.50	58.92	58.50	46.62	69.50	64.53
DT	63.50	60.16	64.50	59.97	60.00	48.69	70.50	68.21	68.00	65.64	64.00	58.12	58.50	47.07	70.00	67.84
MNB	72.50	68.59	68.50	63.12	62.00	51.83	72.00	68.02	73.00	70.31	**69.50**	65.61	**66.50**	**60.10**	**75.00**	**72.72**
GNB	57.50	56.95	65.00	64.96	54.00	56.20	71.50	70.78	57.00	56.51	65.50	**66.00**	54.00	56.22	71.50	70.62
KNN	69.50	67.66	67.00	65.21	55.00	38.29	70.50	67.97	68.50	63.40	56.50	46.43	49.50	30.71	57.00	49.34

Table 5: Accuracy (Acc) and macro-F1 (F1) reported on sentences after English removal. Numbers in bold indicate the best result for each feature.

4.3 Large Language Models

In this experiment, we evaluate multiple bilingual LLMs that were trained on both Arabic and English. We tested the following models: ALLaM-7B-Instruct-preview (Bari et al., 2025), Fanar-1-9B-Instruct (Abbas et al., 2025), AceGPT-v2-32B (Huang et al., 2024), and Jais-2-8B-Chat (Anwar et al., 2025). All models were evaluated on the inference task without fine-tuning across k-shot settings with k=0, 3, 5, and 10. In each k-shot setting, k samples were provided in each class. We use simple and straightforward prompts in Arabic with all models, asking them to analyze the sentiment of a given sentence and choose between positive, negative, or neutral. A sample of a 3-shot prompt is provided in the Appendix. For each prompt setting, experiments were repeated three times, and the average accuracy was reported, as detailed in Table 7.

Among all models, Fanar demonstrates strong performance across settings, achieving 88% accuracy in the 5-shot learning setting, the highest observed result among LLMs. Moreover, Fanar is the only LLM that outperforms our baseline result in 3-, 5-, and 10-shot settings by 1%, 3%, and 0.5%, respectively. The Fanar performance may be attributed to its exposure to dialectal Arabic during training. Following Fanar, ALLaM maintains competitive performance across different settings, achieving the highest accuracy of 83.5% in the 3-shot setting. Notably, Jais-2-8B-Chat outperformed all models in the zero-shot setting, achieving 78.5% accuracy, suggesting an advantage from training on Arabic–English code-switching data alongside Arabic dialects.

All LLMs exhibit performance variations across shots, as shown in Figure 1. Clearly, all models demonstrate performance gains in the 3-shot setting; however, performance declines when the number of shots increases to 5 or 10, with Fanar being the only exception. This reduction in LLMs' performance with an increasing number of shots has been observed in some research on downstream NLP tasks such as Arabic dialect identification (Al-Azani et al., 2024) and sentiment analysis (Zhang et al., 2024). This suggests that increasing the number of shots does not necessarily lead to performance gains.

Model	Accuracy	Macro-F1	Epocs	Batch size	Weight Decay	Learning Rate
Baseline	85.00	84.60	-	-	-	-
SaudiBERT	92.00	91.68	8	5	0.02	3e-5
mBert	88.50	87.92	5	16	0.1	3e-5
XLM-R-Base	**95.50**	**95.53**	5	5	0.01	2e-5
XLM-R-Large	94.00	94.07	10	5	0.01	2e-5
AraBERTv2-Base	90.50	90.27	8	3	0.01	2e-5
AraBERTv2-Large	92.00	91.89	10	32	0.01	5e-5
CAMeLBERT-Da	81.00	79.12	6	5	0.01	2e-5
CAMeLBERT-Mix	83.00	82.04	9	8	0.01	2e-5
GigaBERT-v4-Arabic-and-English	92.50	92.44	5	5	0.02	2e-5

Table 6: Accuracy and macro-F1 for SLMs with optimal configurations.

Model	Zero-shot	3-shot	5-shot	10-shot
ALLaM-7B-Instruct-preview	67.50	83.50	82.00	81.50
Fanar-1-9B-Instruct	60.00	86.00	**88.00**	85.50
AceGPT-v2-32B	48.50	82.50	80.00	63.50
Jais-2-8B-Chat	78.50	82.00	73.50	69.00

Table 7: Accuracy for LLMs across k-shot settings.

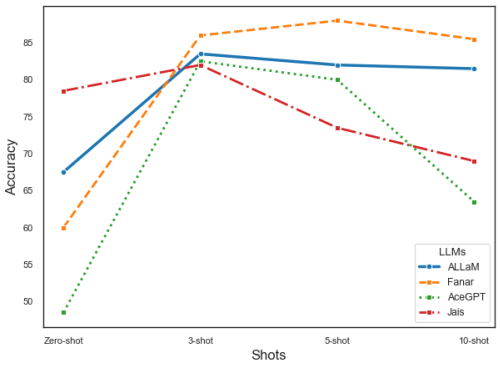

Figure 1: Performance trends of LLMs across different shot settings.

5 Conclusion

In this paper, we conducted the first sentiment analysis study on Saudi-English code-switching texts and introduced the Saudi-English Sentiment Analysis Dataset (SESA-CSD). We established baseline results for this task by applying machine learning models, achieving reasonable and competitive performance. Our baseline was 85% accuracy and 84.60% macro-F1, achieved by a voting classifier combining logistic regression, linear SVM, and Gaussian Naive Bayes. We highlighted that removing English words during the cleaning phase negatively affected the machine learning models, resulting in approximately 9% and 10% reductions in accuracy and macro-F1, respectively. By applying state-of-the-art multilingual and Arabic small language models, such as XLM-ROBERTa, SaudiBERT, and GigaBERT, we improved upon the baseline results by 3% to 11% in both accuracy and macro-F1, respectively. Applying bilingual Arabic-English large language models did not result in improvements over the baseline, with Fanar being the exception, which improved performance by 3% in the 5-shot setting.We conclude that multilingual and Arabic small language models outperform bilingual Arabic-English large language models in zero-shot and few-shot settings for sentiment analysis of Saudi-English code-switching text. Our findings underscore the need for additional datasets to study code-switching phenomena in Saudi-English text and the necessity of training language models on code-switching data to achieve optimal performance.

6 Limitations

The dataset used in this study is synthetic and was generated by GPT-4. Using synthetic data offers a practical solution to many challenges, as collecting code-switching data is non-trivial. Most code-switching occurs among younger generations on social media platforms such as WhatsApp, which requires consent from each participant, a requirement that is often impractical. For other platforms, such as X, obtaining a sufficient number of tweets

that represent code-switching can be costly. Spoken code-switching also occurs in educational or professional settings, which similarly necessitates both participant consent and transcription for analysis. Consequently, the use of synthetic datasets provides a practical solution to these challenges. However, we have noticed that some samples may exaggerate code-switching and adopt stylistic patterns that do not reflect naturally occurring code-switching in the Saudi community. Despite filtering the samples to exclude phrases or words from MSA and other Arabic dialects, we could not filter out samples that do not fully align with naturally occurring code-switching practices in Saudi Arabia. Therefore, experimenting with naturally occurring code-switching data might yield substantially different results. We expected similar results with simple, short sentences, but anticipated a decline with complex sentences that include more English and dialectal words and phrases.

In addition, Saudi Arabia has various dialects, such as Najdi and Hijazi, which were not distinguished in this study and were treated as a single dialect. Moreover, the dataset is relatively small and imbalanced; therefore, conducting the same experiments on larger and/or more balanced datasets could yield different performance.

Acknowledgments

All experiments in this study were conducted using the Aziz Supercomputer at King Abdulaziz University, Jeddah, Saudi Arabia. We would like to express our gratitude to the Aziz team for their support throughout this research.

References

Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, and 22 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pretraining bert on arabic tweets: Practical considerations. *Preprint*, arXiv:2102.10684.

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

Aseel Addawood, Alhanouf Alsuwailem, Ali Alohali, Dalal Alajaji, Mashail Alturki, Jaida Alsuhaibani, and Fawziah Aljabli. 2020. Tracking And Understanding Public Reaction During COVID-19: Saudi Arabia As A Use Case. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Wafia Adouane, Samia Touileb, and Jean-Philippe Bernardy. 2020. Identifying sentiments in Algerian code-switched user-generated comments. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2698–2705, Marseille, France. European Language Resources Association.

Sadam Al-Azani, Nora Alturayeif, Haneen Abouelresh, and Alhanoof Alhunief. 2024. A comprehensive framework and empirical analysis for evaluating large language models in arabic dialect identification. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Hamed Al-Rubaiee, Renxi Qiu, and Dayou Li. 2016. Identifying Mubasher software products through sentiment analysis of Arabic tweets. In *2016 International Conference on Industrial Informatics and Computer Systems (CIICS)*, pages 1–6, Sharjah, Dubai, United Arab Emirates. IEEE.

Sadeen Alharbi, Raghad Aloraini, Reem BinMuqbil, Ahmed Ali, Saiful Bari, Areeb Alowisheq, and Yaser Alonaizan. 2024. Leveraging llm for augmenting textual data in code-switching asr: Arabic as an example. In *Proceedings of the INTERSPEECH 2024 on Synthetic Data's Transformative Role in Foundational Speech Models (SynData4GenAI)*, Kos, Greece. ISCA.

Tasneem S. Almasah, Gamal A. Ebrahim, and Marwa A. Abdelaal. 2023. A code-switched arabic-english sentiment analysis approach based on deep-learning. In *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 452–457.

Latifah Almuqren and Alexandra Cristea. 2021. AraCust: a Saudi Telecom Tweets corpus for sentiment analysis. *PeerJ Computer Science*, 7:e510.

Kais Sultan Mousa Alowidha. 2024. English-arabic code switching and identity in bilingual saudis living in saudi arabia: A comparative study between large and small cities. *Educational Administration: Theory and Practice*, 30(5):4713–4722.

Dhuha Alqahtani, Lama Alzahrani, Maram Bahareth, Nora Alshameri, Hend Al-Khalifa, and Luluh Aldhubayi. 2022. Customer sentiments toward saudi banks during the covid-19 pandemic. In *Proceedings of the 5th International Conference on Natural Language and*

Speech Processing (ICNLSP 2022), pages 251–257, Trento, Italy. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Mohamed Anwar, Abdelhakim Freihat, George Ibrahim, Mostafa Awad, Abdelrahman Atef Mohamed Ali Sadallah, Gurpreet Gosal, Gokul Ramakrishnan, Biswajit Mishra, Sarath Chandran, Ahmed Frikha, Rituraj Joshi, Etienne Goffinet, Abhishek Maiti, Ali El Filali, Sarah Al Barri, Samujjwal Ghosh, Rahul Pal, Parvez Mullah, Awantika Shukla, and 41 others. 2025. Jais 2: A family of Arabic-centric open large language models. Technical report, IFM.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Wassim El-Hajj, Khaled Bashir Shaban, Nizar Habash, Ahmad Al-Sallab, and Ali Hamdi. 2019. A survey of opinion mining in arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(3).

Fazlourrahman Balouchzahi and H L Shashirekha. 2021. LA-SACo: A study of learning approaches for sentiments analysis inCode-mixing texts. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 109–118, Kyiv. Association for Computational Linguistics.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *The Thirteenth International Conference on Learning Representations*.

Afnan Bayazed, Ola Torabah, Redha AlSulami, Dimah Alahmadi, Amal Babour, and Kawther Saeedi. 2020. SDCT: Multi-Dialects Corpus Classification for Saudi Tweets. *International Journal of Advanced Computer Science and Applications*, 11(11).

Naaima Boudad, Rdouan Faizi, and Oulad haj thami Rachid. 2023. Multilingual, monolingual and mono-dialectal transfer learning for moroccan arabic sentiment classification. *Social Network Analysis and Mining*, 14.

Communications,Space and Technology Commission. 2025. Cst issued the saudi internet report 2024. `https://www.cst.gov.sa/ar/media-center/news/N2025051200` [Accessed: 2025-12-25].

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *Commun. ACM*, 64(4):72–81.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. 2025. Darijabert: a step forward in nlp for the written moroccan dialect. *International Journal of Data Science and Analytics*, 20:917–929.

Björn Gambäck and Amitava Das. 2014. On measuring the complexity of code-mixing. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 1–7, Goa, India. 1st Workshop on Language Technologies for Indian Social Media.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *Interspeech 2017*, pages 67–71.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of

variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mohamed Amine Jerbi, Hadhemi Achour, and Emna Souissi. 2019. Sentiment analysis of code-switched tunisian dialect: Exploring rnn-based techniques. In *Arabic Language Processing: From Theory to Practice*, volume 1108 of *Communications in Computer and Information Science*, pages 122–131, Cham. Springer.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Working Paper 43, University of Pennsylvania, Annenberg School for Communication.

Yash Kumar Lal, Vaibhav Kumar, Mrinal Dhar, Manish Shrivastava, and Philipp Koehn. 2019. De-mixing sentiment from code-mixed text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 371–377, Florence, Italy. Association for Computational Linguistics.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An empirical study of pre-trained transformers for Arabic information extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4727–4734, Online. Association for Computational Linguistics.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Springer International Publishing.

Jiaxiang Liu, Xuyi Chen, Shikun Feng, Shuohuan Wang, Xuan Ouyang, Yu Sun, Zhengjie Huang, and Weiyue Su. 2020. Kk2018 at SemEval-2020 task 9: Adversarial training for code-mixing sentiment classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 817–823, Barcelona (online). International Committee for Computational Linguistics.

Yili Ma, Liang Zhao, and Jie Hao. 2020. XLP at SemEval-2020 task 9: Cross-lingual models with focal loss for sentiment analysis of code-mixing language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 975–980, Barcelona (online). International Committee for Computational Linguistics.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching 1. *Linguistics*, 18:581–618.

Faisal Qarah. 2024. Saudibert: A large language model pretrained on saudi dialect corpora. *Preprint*, arXiv:2405.06239.

Gagan Sharma, R Chinmay, and Raksha Sharma. 2023. Late fusion of transformers for sentiment analysis of code-switched data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6485–6490, Singapore. Association for Computational Linguistics.

Ahmed Sherif and Caroline Sabty. 2024. Sentiment analysis for egyptian arabic-english code-switched data using traditional neural models and advanced language models. In *Speech and Computer: 26th International Conference, SPECOM 2024, Belgrade, Serbia, November 25–28, 2024, Proceedings, Part II*, page 54–69, Berlin, Heidelberg. Springer-Verlag.

Sara Stefanich, Jennifer Cabrelli, Dustin Hilderman, and John Archibald. 2019. The morphophonology of intraword codeswitching: Representation and processing. *Frontiers in Communication*, Volume 4 - 2019.

Sergei Ternovykh and Anastasia Nikiforova. 2023. Recent advances in textual code-switching. In *Natural Language Processing and Information Retrieval: Principles and Applications*, chapter 6, pages 159–184. CRC Press, Taylor & Francis Group.

Vision2030. 2025. Strategies custodian of the two holy mosques scholarship program. `https://www.vision2030.gov.sa/en/explore/strategies/custodian-of-the-two-holy-mosques-scholarship-program`[Accessed: 2025-12-29].

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

Figure 2 presents a sample of a 3-shot prompt used to instruct LLMs.

Arabic Prompt	Translation
3-shot Prompt: حلل المشاعر في الجملة التالية. اختر إجابة واحدة فقط من ايجابي أو سلبي أو محايد. **الجملة:** الحفلة كانت amazing أمس. **الإجابة:** إيجابي **الجملة:** اليوم الجو كان really nice بالمرة. **الإجابة:** إيجابي **الجملة:** ترى the new movie مره حلو لازم تشوفه! **الإجابة:** إيجابي **الجملة:** We went to the mall ومع الأسف كان زحمة! **الإجابة:** سلبي **الجملة:** اليوم الشغل كان too stressful بصراحة. **الإجابة:** سلبي **الجملة:** اليوم عندي final exam ومره متوتر! **الإجابة:** سلبي **الجملة:** الويكند الجاي بنسافر للخبرwith the family **الإجابة:** محايد **الجملة:** اليوم حروح مع أصحابي to the mall. **الإجابة:** محايد **الجملة:** أنا رايح للمول today عشان أشتري بعض الأشياء للبيت. **الإجابة:** محايد **الجملة:** I can't believe إنه الجو صار حلو جدا اليوم! **الإجابة:**	**Translated 3-shot Prompt:** Analyze the sentiment of the following sentence. Choose only one answer from positive, negative, or neutral. **Sentence:** The party was amazing yesterday. **Answer:** Positive **Sentence:** Today the weather was really nice, honestly. **Answer:** Positive **Sentence:** The new movie is really good, you have to watch it! **Answer:** Positive **Sentence:** We went to the mall, and unfortunately it was crowded! **Answer:** Negative **Sentence:** Today, work was too stressful, honestly. **Answer:** Negative **Sentence:** Today I have a final exam, and I am really nervous. **Answer:** Negative **Sentence:** Next weekend we will travel to Alkobar with the family. **Answer:** Neutral **Sentence:** Today I am going with my friends to the mall. **Answer:** Neutral **Sentence:** I am going to the mall today to buy some things for the house. **Answer:** Neutral **Sentence:** I can't believe the weather became beautiful today. **Answer:**

Figure 2: Sample of a 3-shot prompt for LLMs.