Evaluating ChatGPT and Bard AI on Arabic Sentiment Analysis

Abdulmohsen Al-Thubaity[1], **Sakhar Alkhereyf**[1], **Hanan Murayshid**[1], **Nouf Alshalawi**[1], **Maha Bin Omirah**[2], **Raghad Alateeq**[2], **Rawabi Almutairi**[2], **Razan Alsuwailem**[3], **Manal Alhassoun**[1], **Imaan Alkhanen**[1]
[1]King Abdulaziz City for Science and Technology, Saudi Arabia
[2]Imam Mohammad ibn Saud University, Saudi Arabia
[3]Qassim University, Saudi Arabia
{aalthubaity,salkhereyf,hmurayshed,nalshalawi}@kacst.edu.sa

Abstract

Large Language Models (LLMs) such as ChatGPT and Bard AI have gained much attention due to their outstanding performance on a range of NLP tasks. These models have demonstrated remarkable proficiency across various languages without the necessity for full supervision. Nevertheless, their performance in low-resource languages and dialects, like Arabic dialects in comparison to English, remains to be investigated. In this paper, we conduct a comprehensive evaluation of three LLMs for Dialectal Arabic Sentiment Analysis: namely, ChatGPT based on GPT-3.5 and GPT-4, and Bard AI. We use a Saudi dialect Twitter dataset to assess their capability in sentiment text classification and generation. For classification, we compare the performance of fully fine-tuned Arabic BERT-based models with the LLMs in few-shot settings. For data generation, we evaluate the quality of the generated new sentiment samples using human and automatic evaluation methods. The experiments reveal that GPT-4 outperforms GPT-3.5 and Bard AI in sentiment analysis classification, rivaling the top-performing fully supervised BERT-based language model. However, in terms of data generation, compared to manually annotated authentic data, these generative models often fall short in producing high-quality Dialectal Arabic text suitable for sentiment analysis.

1 Introduction

Sentiment analysis is the task of determining the emotional tone of a piece of text, such as whether it is positive, negative, or neutral. It is a challenging task for many languages, including Arabic, due to the complex morphology and syntax of the language. Various approaches have been used to tackle this challenge, including rule-based and dictionary-based methods (ElSahar and El-Beltagy, 2014; Al-Twairesh et al., 2016; Al-Thubaity et al., 2018b), classical machine learning algorithms (Abdul-Mageed et al., 2014; Duwairi and Qarqaz, 2014; Abdulla et al., 2013; Mourad and Darwish, 2013; Abdul-Mageed et al., 2011), deep learning (Alayba et al., 2018), and pre-trained language models such as BERT (Devlin et al., 2018).

However, sentiment analysis faces a critical challenge, particularly in the context of social media, which is data drift and concept drift (Zhao et al., 2022). This challenge necessitates continuous monitoring of sentiment analysis models and updating rule-based systems and dictionaries, if utilized, as well as retraining machine learning models with new data.

Recent advancements in NLP, particularly the emergence of large language models (LLM) such as GPT-4 (OpenAI, 2023) and PaLM 2 (Anil et al., 2023), and their utilization in ChatGPT and Bard AI, respectively, show potential in countering the issues of data and concept drift in sentiment analysis. These LLMs are trained on large and diverse datasets and fine-tuned or prompted for various tasks, including sentiment analysis (Wang et al., 2023; Qin et al., 2023; Kocoń et al., 2023; Amin et al., 2023), and have proven their capabilities for this task in English.

Although few research efforts have been made to test the ability of LLMs for Arabic sentiment analysis, focusing on single language models like AraT5 (Elmadany et al., 2022) or multiple models, including ChatGPT and others (Khondaker et al., 2023), to the best of our knowledge, no study has been conducted to evaluate both of Bard AI and ChatGPT LLMs for Arabic Sentiment Analysis. In particular, this is the first attempt to evaluate Bard AI on Arabic Sentiment Analysis.

This paper aims to evaluate three generative large language models, namely Generative Pre-trained Transformers GPT-3.5 and GPT-4 through ChatGPT by OpenAI, and the Pathways Language Model (PaLM) through Bard by Google on a sentiment analysis dataset comprising in the Saudi di-

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 335–349

alect, the Saudi Dialect Twitter Corpus (SDTC) (Al-Thubaity et al., 2018a), comprising 5,400 tweets classified into five classes: positive, negative, neutral, spam, and "I do not know" class, to reveal the capabilities of LLMs in tackling the Arabic sentiment analysis challenge.

The contribution of this paper is twofold. First, it includes the evaluation of Google Bard AI for the first time in this type of analysis. Second, it evaluates these models for Arabic sentiment analysis from different and novel perspectives, as illustrated in the experiment's design (section 3). Mainly, we address the following Research Questions (RQs):

- RQ1: How is the performance of generative models when compared with fully supervised models in a relatively challenging and subjective task for Arabic NLP, namely, Arabic Sentiment Analysis? We investigate when there are few or no available training examples for generative models and compare the performance with fully fine-tuned BERT-based models.
- RQ2: What is the difference in the performance of widely used generative models on the Arabic Sentiment Analysis? In particular, we use ChatGPT (both GPT 3.5 and GPT-4) and Bard AI (PaLM 2), which, to the best of our knowledge, is the first paper to evaluate Bard AI on Arabic Sentiment Analysis.
- RQ3: How good are these models for generating new sentiment data examples in Arabic dialects? We investigate this in two ways: 1) manual evaluation of the generated examples. 2) using these examples as training samples for BERT-based models and comparing the performance with manually annotated data.

The structure of the rest of the paper is as follows: Section 2 presents previous work on sentiment analysis and using generative models for natural understanding tasks. Section 3 shows the experiment design and the dataset used to evaluate various models. Then, in section 4, we present the results of four comprehensive experiments and analysis. We conclude the paper in section 5.

2 Related Work

2.1 Arabic Sentiment Analysis Corpora

Over the past ten years, Sentiment Analysis research, especially in the Arabic language, has gained significant interest because of the accessible sentiment data primarily from social media platforms like Twitter. The growth of social media has enabled Arabic speakers to write in their dialects, which was previously limited to the spoken form due to the language's diglossic nature. This has resulted in an abundance of dialectal textual data without the formality of standards, unlike Modern Standard Arabic (MSA) (Darwish et al., 2021). Numerous datasets have emerged for Arabic sentiment analysis across different genres, mainly tweets, with a majority in Arabic dialects, including Egyptian (Nabil et al., 2015; Refaee and Rieser, 2014), Levantine (Baly et al., 2018), Maghrebi (Mdhaffar et al., 2017; Zarra et al., 2017), and Saudi Dialect (Al-Thubaity et al., 2018a; Al-Twairesh et al., 2017; Assiri et al., 2016), among others. Other datasets cover multiple Arabic dialects in addition to Modern Standard Arabic (MSA) (Elmadany et al., 2018; Al-Obaidi and Samawi, 2016; Abdul-Mageed et al., 2014).

2.2 Arabic Sentiment Analysis Methods

Historically, much like other languages, Arabic Sentiment Analysis relied on rule-based methods, focusing primarily on crafting sentiment lexicons (ElSahar and El-Beltagy, 2014; Al-Twairesh et al., 2016; Al-Thubaity et al., 2018b). In more recent years, there has been a growing interest in using machine learning methods for Arabic Sentiment Analysis. These methods can learn the patterns of sentiment from a large corpus of text, and they are not as susceptible to the limitations of lexicon-based methods. Notable machine learning techniques employed include Naïve Bayes (NB), Support Vector Machines (SVMs), and K-Nearest Neighbor (k-NN) classifiers, leveraging morphological and syntactic features (Abdul-Mageed et al., 2014; Duwairi and Qarqaz, 2014; Abdulla et al., 2013; Mourad and Darwish, 2013; Abdul-Mageed et al., 2011).

The rapid evolution of natural language processing (NLP) has been marked by the introduction and success of transformer-based models, particularly BERT (Devlin et al., 2018) in 2018. Following that, other transformer-based models for natural language understanding (NLU) have been proposed, such as RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and ALBERT (Lan et al., 2020). Many of these models were pre-trained on mono-lingual datasets, mainly in English. Also, multilingual models were released, such as mBERT (Devlin

et al., 2018), or language-specific models (other than English). Remarkably, there have been proposed Arabic-specific pre-trained language models, for example, ArBERT (Abdul-Mageed et al., 2021), MarBERT (Abdul-Mageed et al., 2021), AraBERT (Antoun et al., 2020), and CAMEL-BERT (Inoue et al., 2021). BERT and BERT-like models achieved state-of-the-art performance on many NLP tasks, including sentiment analysis in many languages Sun et al. (2019).

2.3 Generative Models for Arabic NLP

While BERT and BERT-like models are discriminative models for NLU tasks, the NLP community also witnessed a surge in the development and application of generative models designed to produce new text samples. Examples of generative models include, the GPT (Radford et al., 2018, 2019; Brown et al., 2020), T5 (Raffel et al., 2020), and BLOOM (Scao et al., 2022). Similar to BERT-like models, there have been proposed multilingual and language-specific generative models and, more specifically for Arabic, such as AraT5 (Elmadany et al., 2022), and AraGPT-2 (Antoun et al., 2021). Generative models have shown promise in tasks like text completion, translation, summarization, and even sentiment analysis, where they can be used to generate sentiment-consistent text expansions, modifications, or new text examples. In particular, Elmadany et al. (2022) shows that the AraT5 model outperforms state-of-the-art models on several Arabic language generation tasks. AraT5 is pre-trained on a large Arabic text and code dataset and fine-tuned on diverse Arabic language generation tasks, including machine translation, summarization, question answering, and paraphrasing.

2.4 Evaluating Generative LLMs

The introduction of generative Large Language Models (LLMs) like ChatGPT and Bard AI marked a significant milestone in the journey of generative models. These models, built on more advanced versions of the transformer architecture, such as GPT-3, GPT-4, and PaLM, demonstrated human-like text generation capabilities in multiple languages, including Arabic. Following this trend, there has been a growing interest in evaluating the capabilities of generative models, mainly ChatGPT and Google Bard AI, for various NLP tasks, such as sentiment analysis (Wang et al., 2023; Qin et al., 2023; Kocoń et al., 2023; Amin et al., 2023), summarization (Qin et al., 2023; Alyafeai et al., 2023; Khondaker et al., 2023), and POS tagging (Alyafeai et al., 2023; Abdelali et al., 2023). Initial methodological efforts to evaluate these models focus on their performance in high-resource languages such as English (Qin et al., 2023; Bubeck et al., 2023). Other studies have evaluated LLMs for their performance on other low-resource languages (Ahuja et al., 2023; Bang et al., 2023; Lai et al., 2023). The findings from these studies indicate that the trending ChatGPT is a capable language model, but it does not surpass the current state-of-the-art (SOTA) solutions in most NLP tasks. However, when it comes to sentiment analysis, which is the main focus of our research, one study (Amin et al., 2023) contradicted the majority and found that ChatGPT outperformed the leading solution, suggesting that this is a promising area for further research. Furthermore, most studies on sentiment analysis using ChatGPT have been conducted on English datasets, and a few research in the Arabic language. Therefore, our research aims to bridge the gap in sentiment analysis research for the Arabic language and demonstrate the potential of ChatGPT in understanding and analyzing sentiment in this context.

For Arabic, Khondaker et al. (2023) present a comprehensive evaluation of ChatGPT's performance on 32 Arabic NLP tasks, including sentiment analysis. The results suggest that, although ChatGPT performs satisfactorily on most Arabic NLP tasks, it is consistently surpassed by the smaller Arabic-focused, fully supervised, fine-tuned model, AraT5. Alyafeai et al. (2023) investigate the performance of the two ChatGPT models, GPT-3.5 and GPT-4, on seven Arabic NLP tasks and compare their performance against SoTA models. On the sentiment analysis task, the results show that GPT-4 outperforms GPT-3.5. However, both models are outperformed by the SoTA model, i.e., MARBERT (Abdul-Mageed et al., 2021). They show that GPT-4 was more robust to different prompts, and its performance improved with the increase in the number of few-shot examples, unlike GPT-3.5. Another study (Abdelali et al., 2023) demonstrated that the SoTA model (with 0.760 F-1 score) outperformed ChatGPT (with 0.550 F1 score) on an Arabic sentiment analysis dataset. However, the ChatGPT model was only evaluated in a zero-shot learning setting, meaning that it was not given any example of the sentiment analysis task. We notice that the only study that includes

Bard AI for Arabic NLP tasks is (Kadaoui et al., 2023), which conducted a comprehensive assessment of Bard AI and ChatGPT, covering both GPT-3.5 and GPT-4 in the domain of machine translation across ten varieties of Arabic.

In contrast to the studies mentioned above, we also include Google's Bard AI in our evaluation for Arabic Sentiment Analysis. This is significant because, to the best of our knowledge, although there are some studies that use ChatGPT for Arabic Sentiment Analysis, no other comparable research has been conducted to evaluate both Bard AI and ChatGPT on Arabic Sentiment Analysis.

3 Experiments Design and Data

The primary objective of these experiments is to assess the capabilities of generative models for Arabic sentiment analysis and the potential of data augmentation and generation for this task. We evaluate three models:

- Generative Pre-trained Transformers GPT-3.5,
- and GPT-4, both accessed via ChatGPT by OpenAI.
- PaLM 2 facilitated by Bard AI by Google. Throughout the paper, the terms "Bard" and "PaLM" will be used interchangeably.

For GPT-3.5 and GPT-4 we utilize the ChatGPT API to send prompts and receive responses. For PaLM 2, prompts are manually sent to Bard AI via its web interface, from which we extract the relevant responses. Also, we utilize these models to generate new samples and systematically evaluate the generated examples.

We have four main experiments:

- **Exp.1**: As a baseline, we train various Arabic BERT-based language models using an existing dataset and assess their performance on the dataset. We utilize models pre-trained on various Arabic corpora, specifically those trained on Twitter or Arabic dialectal data. The model that shows the best performance will serve as our baseline model.
- **Exp.2**: We evaluate the performance of the generative models (i.e., GPT-3.5, GPT-4, and PaLM 2) by instructing them to classify the test data into positive, negative, or neutral categories. We conduct the evaluation using k shots. We will assess the performance of each model against the test data and compare it to the best model identified in Exp. 1. This experiment aims to address RQ1 and RQ2.
- **Exp.3**: We prompt the generative models with a given sentiment (positive, negative, and neutral), instructing them to generate m tweets. Samples of the generated tweets will be manually evaluated for their naturalness using various criteria. This experiment aims to address RQ3.
- **Exp.4**: The data generated in Exp.3 will be utilized in two different ways. Firstly, it will be used to augment the original training data, which will then be fine-tuned with the BERT-based models used in Exp 1. Secondly, the synthesized data will be used to fine-tune the BERT-based models used in Exp 1. For both approaches, the performance will be evaluated against the test data. This experiment aims to address RQ3.

For the abovementioned experiments, we use the Saudi Dialect Twitter Corpus (SDTC) (Al-Thubaity et al., 2018a). SDTC comprises 5,400 tweets distributed across five classes: positive, negative, neutral, spam, information, and difficult to classify. In our experiments, we focused on the first three classes: positive, negative, and neutral tweets, amounting to 558, 1,632, and 500, respectively. The total number of tweets in our experiments is 2,690.

We randomly split the SDTC dataset into 75% for training $\text{SDTC}_{\text{train}}$ and 25% for testing $\text{SDTC}_{\text{test}}$, obeying the class distribution. Also, we selected 30 tweets from each class (90 tweets overall) from $\text{SDTC}_{\text{train}}$ to evaluate the output of each proposed prompt. We use SDTC_{dev} to refer to these 90 tweets.

$\text{SDTC}_{\text{train}}$ is used to fine-tune the language models in Exp 1 and Exp 4. The $\text{SDTC}_{\text{test}}$ set is used for evaluating the fine-tuned language models (Exp 1 and Exp 4) and for the predictions of the generative models in Exp 2. Experiment 3 involves human judgment, and the outputs of the generative models will be used to fine-tune the language models in Exp 4. We use SDTC_{dev} for evaluating the output of different prompts in Exp 2 and Exp 3. We make SDTC_{dev} balanced in classes because of its relatively small size, due to budget and time constraints, as we couldn't evaluate all prompts and different numbers of shots on a larger scale.

However, we evaluate the best settings in terms of prompts and number of shots on the whole test set $SDTC_{test}$. These prompts were inspired or adapted from previously published research (Alyafeai et al., 2023; Khondaker et al., 2023).

4 Experiments and Results

In this section, we show and discuss the results of the experiments described in section 3. To assess the performance of the models, we employed the accuracy (Acc) metric, along with the micro-averages of precision (P), recall (R), and F-1 score (F) values. When evaluating the models' performance, we focus on the F1 measure as the primary metric of comparison.

4.1 Experiment 1: Fine-tuning BERT Models (baseline)

We fine-tuned five Arabic BERT-based models using the training data, $SDTC_{train}$, and evaluated their performance on the test data, $SDTC_{test}$. Namely, we fine-tune:

- bert-large-arabertv02-twitter and bert-base-arabertv02-twitter (Antoun et al., 2020).
- MARBERTv2 (Abdul-Mageed et al., 2021).
- bert-base-arabic-camelbert-da (Inoue et al., 2021).
- and bert-base-qarib (Abdelali et al., 2021).

Model	Acc	P	R	F-1
arabert-base	**79**	**79**	**79**	**79**
arabert-large	78	77	78	77
qarib	78	77	77	77
MARBERT	77	76	77	76
camelbert	72	72	72	72

Table 1: Performance measures for the five fine-tuned language models. We show the micro-averaged score for each metric.

These models were fully or partially pre-trained on Twitter or Arabic dialect data. Numerous experiments were conducted across all models, involving varying hyperparameter values. Appendix B shows the details of hyperparameters and experimental setups. Table 1 shows the results of fine-tuning the five models.

The results demonstrate that models solely pre-trained on the same data as the fine-tuning data exhibit the best performance, in our case, Twitter data. Notably, the performance of the bert-base-arabertv02-twitter model outperforms the larger bert-large-arabertv02-twitter model, contrary to the typical expectation.

For further analysis, see Appendix C, where we show that the best BERT-based, i.e., bert-base-arabertv02-twitter, has the highest confusion when differentiating between positive and neutral classes.

4.2 Experiment 2: Sentiment Analysis with Generative Models

In this set of experiments, we evaluate ChatGPT (GPT 3.5 and GPT 4) and Bard AI on $SDTC_{test}$. Unlike the setup of hyperparameters for pre-trained language models, which are known and controlled, determining the optimal prompt design for generative models involves trial and error processes. We conducted experiments with seven prompt designs in Arabic and English to classify tweets in $SDTC_{test}$. We evaluate each prompt design on $SDTC_{dev}$ and then select the prompt with the highest accuracy using k shots where $k = \{0, 1, 3, 5\}$. Each shot is a triplet of a positive, negative, and neutral tweet.

The optimal prompt for Bard AI achieved an accuracy of 0.7 for $k = 5$ (15 tweets overall). It is as follows:

> Given the examples:
>
> positive train tweet ; Sentiment: 1 (positive)
> neutral train tweet ; Sentiment: 0 (neutral)
> negative train tweet ; Sentiment: -1 (negative)
>
> positive train tweet ; Sentiment: 1 (positive)
> neutral train tweet ; Sentiment: 0 (neutral)
> negative train tweet ; Sentiment: -1 (negative)
>
> ...
>
> You are a helpful assistant that can predict whether a given tweet in Arabic is Positive, Negative, or Neutral. Do not show any warning, explanation or disclaimer. Please provide your response for testing tweet in tabular format showing the tweet and the classification.
>
> Testing tweet: Test tweet

For GPT-3.5 and GPT-4, the optimal prompt achieved accuracy scores of 0.81 and 0.91, respectively, for $k = 0$. It is as follows:

> What is the sentiment of the following tweets? Answer with positive, negative, or neutral.
>
> Test tweet

After selecting the best prompt, we evaluate each of the three generative models on $\text{SDTC}_{\text{test}}$, asking them to classify each example as positive, negative, or neutral. If a model declines to classify a tweet due to its unacceptable content for any reason, we set the prediction to be negative. We compare the outcomes of the three generative models with the test data labels and compute the four performance measures. Table 2 shows the performance measures for the three generative models.

Model	Acc	P	R	F-1
GPT-4	75	**82**	75	**77**
Bard AI	**79**	78	**79**	76
GPT-3.5	70	72	70	70
Best BERT	79	79	79	79

Table 2: The performance measures for the three generative models on $\text{SDTC}_{\text{test}}$. We show the micro-averaged score for each metric.

Based on the F-1 score as a reference performance measure, the results show that GPT-4 and Bard AI achieve comparable performance in few-shot settings with the fully supervised BERT-based models. In particular, GPT-4 has a very close performance to the second-best BERT model (i.e., bert-large-arabertv02-twitter) with an F-1 score of 0.77, and it outperforms the other fine-tuned models. Bard AI comes in second with a score of 0.76, which performs relatively well for sentiment analysis classification compared to fully supervised models. Notably, it outperforms one of the BERT-based models and achieves comparable results to the fine-tuned MARBERTv2 model with an F1 score of 0.76. However, GPT-3.5 has low performance, falling behind BERT-based models. The significant difference between the models' performance on the development set SDTC_{dev} and the test $\text{SDTC}_{\text{test}}$ can be attributed to the different class distributions. In particular, Bard AI performs very low in the neutral class, which represents the third of tweets in SDTC_{dev}.

Class	Best BERT model	GPT-4
Negative	**89**	84
Positive	75	**79**
Neutral	54	**58**

Table 3: F-1 scores for each sentiment class for fine-tuned bert-base-arabertv02-twitter and GPT-4 models.

While GPT-4, in a zero-shot setting, has comparable results to the fine-tuned BERT model, their performance for each class varies considerably. Table 3 shows the F1 score for each sentiment class for both models. The results show that both models (BERT and GPT-4) performed best for the classification of negative tweets, followed by positive tweets, and the most difficult classification task was for neutral tweets. The best fine-tuned BERT model (i.e., bert-base-arabertv02-twitter) outperformed GPT-4 for the classification of negative tweets. However, the latter considerably outperformed the former for the classification of positive and neutral tweets, with a 4-point increase in F1 score for both positive and neutral tweets. Again, as for fine-tuned BERT models, the greatest challenge that generative models may face is differentiating between positive and neutral tweets.

4.3 Experiment 3: Data Generation by Generative Models

We instructed each generative model in a zero-shot setting to generate positive, negative, and neutral tweets. We conducted experiments using 11 different prompt designs in both Arabic and English, and then we selected the best prompt based on the evaluation of the resulting output using three criteria:

- The naturalness of the tweets.
- Tweets are in the Saudi dialect.
- Avoidance of overly brief tweets.

We generate multiple outputs for the same prompt and evaluate it on a small scale (a few runs for each prompt), and then we select the best prompt according to the criteria mentioned above. For all generative models, the best prompt was as follows:

Your role is a data engineer who wants to create synthesized examples of tweets for Arabic sentiment analysis. Generate examples of tweets with sentiment classes ["positive", "negative", "neutral"]. Generate 10 examples in {*sentiment*} in the Saudi Dialect; such that each tweet is in a single row in tabular format. You must generate long tweets.

Using the best prompt above, we instructed each generative model (i.e., GPT-3.5, GPT-4, and Bard AI) to generate tweets for each class (i.e., ["positive", "negative", "neutral"]), matching their respective distribution in the training dataset $SDTC_{train}$, i.e., 391, 1,243, and 351 for positive, negative, and neutral tweets, respectively. See Appendix D for examples of the generated tweets.

To assess the quality of the generated tweets and their associated sentiments produced by the generative models, we randomly select 50 tweets from each class for every generative model (a total of 150 tweets per model). Subsequently, two annotators were involved in addressing the following binary inquiries (Yes/No) for each generated tweet:

- Q1 (Making sense): Is the generated tweet linguistically correct and understandable?
- Q2: (Appropriateness for Twitter): Do you expect to see such text on Twitter?
- Q3: (Matching label): Does the generated tweet match the instructed sentiment?

Model	Class	Q1	Q2	Q3	Q1+Q2+Q3
Bard AI	Pos	94	34	98	32
	Neg	**100**	**80**	94	**76**
	Neu	**90**	**72**	50	**30**
	ALL	**95**	**62**	81	**46**
GPT-3.5	Pos	**98**	**78**	98	**76**
	Neg	56	40	94	34
	Neu	74	54	28	20
	ALL	76	57	74	44
GPT-4	Pos	86	58	**100**	52
	Neg	84	46	**100**	40
	Neu	72	56	**52**	28
	ALL	81	53	**84**	40

Table 4: Percentage of affirmative responses for each question and class across the three generative models. Pos: Positive, Neu: Neutral, Neg: Negative, ALL: all classes.

A generated tweet is considered valid for each question if both annotators concur with a "Yes" response; otherwise, the tweet is regarded as invalid for that specific question.

Table 4 demonstrates the percentage of affirmative responses for each question and class across the three generative models. Table 8 in Appendix D showcases examples where the two annotators answered each question with "No".

ALL: The data suggests that Bard AI slightly outperforms GPT-3.5 and GPT-4 when considering all evaluation questions together (46%) or individually, achieving percentages of 95%, 62%, and 81% for Q1, Q2, and Q3, respectively. All models perform well regarding linguistic correctness (Q1) and matching the instructed sentiment (Q3) for Positive and Negative tweets. However, there are challenges with generating tweets that exhibit appropriateness for Twitter (Q2).

Q1: For linguistic correctness and understandability (Q1), Bard AI consistently achieves high percentages across all classes, followed by GPT-4, which performs lower in Neutral tweets. GPT-3.5 has the lowest performance for Negative tweets. This may be attributed to the stricter constraints that prevent it from generating negative content more than Bard AI. In particular, ChatGPT (based on GPT-3.5 and GPT-4) tends to generate nonsense text in dialectal Arabic more than Bard AI.

Q2: Regarding generating tweets that exhibit appropriateness for Twitter (Q2), Bard AI achieved the highest score, specifically for Negative tweets, followed by GPT-3.5 and GPT-4, respectively. However, the latter models demonstrate low scores for Negative tweets.

Q3: Regarding matching the instructed sentiment (Q3), GPT-4 outperforms both Bard AI and GPT-3.5, with a score of 84%, achieving a perfect score of 100% for Positive and Negative tweets. However, its performance in generating Neutral tweets is relatively low (52%). The performance on Q3 for Neutral tweets is also low for the other two models.

Analysis of the generated neutral tweets: To analyze the confusion of neutral tweets with other classes discussed in Exp.1, we compare the neutral tweets generated by the generative models with the labels given by annotators and found that for Bard AI, 96% of the tweets were classified by annotators as positive. For GPT-3.5, 32% were classified as negative and 68% as positive; for GPT-4, 13% were classified as negative and 87% as positive. It

seems that the generative models find it difficult to clearly distinguish neutral content from other types of content, particularly positive content. This is also demonstrated in Exp.1 when we fine-tuned BERT models, where BERT-based models struggle the most with neutral tweets and misclassify them as positive or negative tweets.

4.4 Experiment 4: Fine-tuning Best BERT Model on Generated Data

Data	Acc	P	R	F
SDTC	**79**	**79**	**79**	**79**
Bard AI	69	70	69	67
GPT-3.5	60	64	60	54
GPT-4	68	74	68	66
Bard AI+SDTC	**79**	**79**	**79**	78
GPT-3.5+SDTC	77	77	77	76
GPT-4+SDTC	**79**	**79**	**79**	**79**
All Data	76	77	76	76

Table 5: Performance measures for the using of different data sets on fine-tuning bert-base-arabertv02-twitter model. All Data: Brad AI + GPT-3.5 + GPT-4 + SDTC. For SDTC, we use only the train set, $SDTC_{train}$

We conducted seven experiments using the best-performing BERT model, namely bert-base-arabertv02-twitter, with the same hyperparameters using both the generated data and $SDTC_{train}$ for these experiments. Table 5 shows the performance of fine-tuning bert-base-arabertv02-twitter using the new data. Similar to the previous experiments, our primary focus was on the F1 measure as the key metric for comparison.

The data suggests that, for each generated dataset, the model fine-tuned on Bard AI data demonstrated the best performance on the testing data with an F1 score of 0.67. It was closely followed by the model fine-tuned on data generated by GPT-4, achieving an F1 score of 0.66. The performance data shows a positive correlation with the human evaluation of the generated data in Experiment 3. The relatively lower performance of these models can be attributed to both the fact that the testing data were sampled from a different population and the quality of the classification of the generated data by the generative models.

Combining the original training data with the generated data from each model did not improve the performance of the fine-tuned model. In fact, it might have even led to a decrease in the model's performance. Moreover, combining all the generated data with the original data has a negative impact on performance. This decrease in performance can also be attributed to the same reasons mentioned earlier.

The performance could be improved by using one or more shots of training data to generate new samples using generative models, ensuring higher similarity between the generated data and the original data distribution. Due to time and budget constraints, we could not do so.

5 Conclusion

In this paper, we presented various experiments using three generative models for Arabic Sentiment Analysis in the Saudi dialect: GPT 3.5, GPT-4, and Bard AI (PaLM 2). We compare their performance with fully supervised BERT-based models. We also evaluate the quality of generated examples by these LLMs using manual and automatic methods.

The experiments show that the generative large language models, with little or no training data in few-shot settings, perform relatively well on Arabic Sentiment Analysis compared to fully fine-tuned models. For sentiment analysis text classification, the experiments show that GPT-4 outperforms most of the BERT-based model and is on a par with the second-best BERT-based model. Bard AI comes next with a performance comparable to fully fine-tuned models, while GPT3.5 significantly underperforms the two models and is lower than the BERT-based models. For sentiment generation, all models struggle to generate high-quality text for sentiment analysis in the Saudi Dialect, especially for neutral text. Interestingly, ChatGPT (both GPT-3.5 and GPT-4) tends to generate nonsense text in the Saudi dialect more than Bard AI. Also, implementing safeguards to prevent the generation of harmful or toxic content is crucial for responsible and safe utilization. However, these restrictions can sometimes act as barriers when generating representative text that has a negative sentiment, especially in applications that require a comprehensive representation of human emotions and viewpoints.

Future research should consider comparing the generative models performance among different Arabic dialects and datasets. Also, another direction for future work is analyzing the performance of generative models pre-trained on dialectal Arabic text (such as AraT5) and fully fine-tuned on generating tweets.

Limitations

Due to constraints in time and resources, we need to highlight the following limitations:

- In Experiment 2, which involves sentiment analysis classification, we evaluated the proposed prompts on a limited number of tweets. Particularly, $SDTC_{dev}$ consists of 30 tweets from each class sampled from the training set. Employing more samples for evaluation could lead to identifying better prompts.
- In Experiment 3, focused on data generation, we only conducted experiments in zero-shot settings due to budget and time restrictions. Conducting experiments with a broader number of shots might unveil more robustly generated data. Additionally, the generated tweets were evaluated on a small data sample, utilizing a straightforward binary classification approach across three aspects of the tweets. A more comprehensive evaluation involving larger samples of generated tweets and encompassing a broader array of aspects would provide a more solid assessment.
- In Experiment 4, we solely assessed the performance of the generated data using the best-performing BERT model, namely bert-base-arabertv02-twitter. Undertaking further experimentation with other BERT-based models would yield valuable insights into the performance of these models.

Also, in utilizing large language models (LLMs) like ChatGPT and BARD AI for classifying public datasets, it is important to acknowledge potential limitations tied to data leakage. Given that these models have been trained on vast amounts of data, there is a chance they might have been exposed to, or "seen", some parts of these public datasets, including SDTC, as "pre-training data exposure". Nevertheless, the influence of this exposure might be minor for a few reasons. First, any specific dataset would only be a drop in the ocean, being among billions of tokens on which the models were trained. Second, many public datasets don't have a raw text structure, reducing direct familiarity, which is the case with SDTC. Lastly, we have shown in our experiments that both ChatGPT and Bard AI don't have perfect results on SDTC, further suggesting that any prior exposure may not significantly skew outcomes.

Ethics Statement

The results obtained in this study must be considered within the framework of the intended usage of the generative models and the criteria applied for their evaluation. The disparities in performance observed among the models could potentially stem from variances in training data, model architecture, or prompt design. Further analysis and exploration will contribute to identifying the underlying causes of these discrepancies. Additionally, our study does not address biases within the models or their approach to handling Arabic content, whether generated directly by the models or translated from other languages. The findings of this study cannot be universally extrapolated to other tasks or various Arabic dialects without undergoing comprehensive investigations.

Acknowledgement

The authors would like to thank the anonymous reviewers for their valuable comments and feedback.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training BERT on Arabic tweets: Practical considerations.

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, et al. 2023. Benchmarking Arabic AI with large language models. *arXiv preprint arXiv:2305.14982*.

Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28(1):20–37.

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

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In

2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT), pages 1–6. IEEE.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative AI. *arXiv preprint arXiv:2303.12528*.

Ahmed Y Al-Obaidi and Venus W Samawi. 2016. Opinion mining: Analysis of comments written in Arabic colloquial. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1.

Abdulmohsen Al-Thubaity, Mohammed Alharbi, Saif Alqahtani, and Abdulrahman Aljandal. 2018a. A Saudi dialect Twitter corpus for sentiment and emotion analysis. In *2018 21st Saudi computer society national computer conference (NCC)*, pages 1–6. IEEE.

Abdulmohsen Al-Thubaity, Qubayl Alqahtani, and Abdulaziz Aljandal. 2018b. Sentiment lexicon for sentiment analysis of Saudi dialect tweets. *Procedia computer science*, 142:301–307.

Nora Al-Twairesh, Hend Al-Khalifa, and AbdulMalik Al-Salman. 2016. AraSenTi: Large-scale Twitter-specific Arabic sentiment lexicons. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 697–705.

Nora Al-Twairesh, Hend Al-Khalifa, AbdulMalik Al-Salman, and Yousef Al-Ohali. 2017. AraSenTi-tweet: A corpus for Arabic sentiment analysis of Saudi tweets. *Procedia Computer Science*, 117:63–72.

Abdulaziz M Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. 2018. A combined CNN and LSTM model for Arabic sentiment analysis. In *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2*, pages 179–191. Springer.

Zaid Alyafeai, Maged S Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. 2023. Taqyim: Evaluating Arabic NLP tasks using chatgpt models. *arXiv preprint arXiv:2306.16322*.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general AI? a first evaluation on ChatGPT. *arXiv preprint arXiv:2303.03186*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 technical report.

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207.

Adel Assiri, Ahmed Emam, and Hmood Al-Dossari. 2016. Saudi Twitter corpus for sentiment analysis. *International Journal of Computer and Information Engineering*, 10(2):272–275.

Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. 2018. ArSentD-LEV: A multi-topic corpus for target-based sentiment analysis in Arabic Levantine tweets. In *OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*, page 37.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

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, et al. 2021. A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, 64(4):72–81.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rehab M Duwairi and Islam Qarqaz. 2014. Arabic sentiment analysis using supervised classification. In *2014 International Conference on Future Internet of Things and Cloud*, pages 579–583. IEEE.

A Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. Arsas: An Arabic speech-act and sentiment corpus of tweets. *OSACT*, 3:20.

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

Hady ElSahar and Samhaa R El-Beltagy. 2014. A fully automated approach for Arabic slang lexicon extraction from microblogs. In *International conference on intelligent text processing and computational linguistics*, pages 79–91. Springer.

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 *6th Arabic Natural Language Processing Workshop, WANLP 2021*, pages 92–104. Association for Computational Linguistics (ACL).

Karima Kadaoui, Samar M Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. TARJAMAT: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. *arXiv preprint arXiv:2308.03051*.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. *arXiv preprint arXiv:2305.14976*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, page 101861.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Salima Mdhaffar, Fethi Bougares, Yannick Esteve, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Third Arabic Natural Language Processing Workshop (WANLP)*, pages 55–61.

Ahmed Mourad and Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 55–64.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2515–2519.

OpenAI. 2023. GPT-4 technical report.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Eshrag Refaee and Verena Rieser. 2014. An Arabic Twitter corpus for subjectivity and sentiment analysis. In *9th International Language Resources and Evaluation Conference*, pages 2268–2273. European Language Resources Association.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is ChatGPT a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Taoufiq Zarra, Raddouane Chiheb, Rajae Moumen, Rdouan Faizi, and Abdellatif El Afia. 2017. Topic and sentiment model applied to the colloquial Arabic: a case study of Maghrebi Arabic. In *Proceedings of the 2017 international conference on smart digital environment*, pages 174–181.

Zhixue Zhao, George Chrysostomou, Kalina Bontcheva, and Nikolaos Aletras. 2022. On the impact of temporal concept drift on model explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4039–4054, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A SDTC Statistics

Table 6 illustrates the statistics of SDTC used in our experiments, including three examples from each class (positive, negative, neutral).

B Experimental Setup for Exp.1

For experiment 1 in subsection 4.1, the implementation of all the BERT-based models was carried out using Python 3.9. Fine-tuning experiments were conducted using Tesla GPUs. The following experimental setup was standardized for all the models:

- We utilized the transformers v4.21.1, AutoTokenizer, and Bert For Sequence Classification libraries from Huggingface.
- Optimization was performed using AdamW with a learning rate of 1e-5.
- The number of epochs was set to 15.
- The value for max_grad_norm was set to 1.0.
- The maximum sentence length was constrained to 70 tokens.
- A batch size of 128 was employed.

C Analysis of BERT-based model predictions

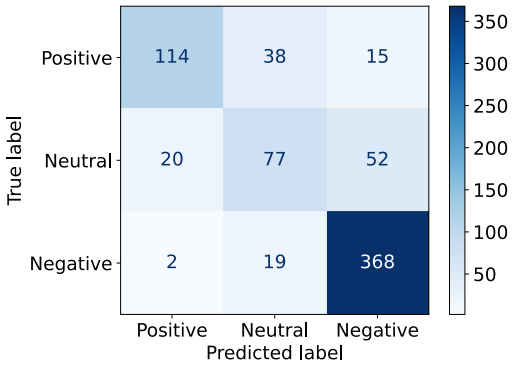

Figure 1: Confusion matrix for the best BERT-based model in Exp.1, bert-base-arabertv02-twitter.

Figure 1 illustrates the confusion matrix for the best-performing model, bert-base-arabertv02-twitter discussed in Exp.1 (subsection 4.1). Overall, the model demonstrates strong performance, showcasing high accuracy in predicting the Negative class (95%) while achieving lower accuracy in predicting the Positive class (68%) and the Neutral class (52%).

The data indicates that the model encounters the greatest challenges when differentiating between Positive and Neutral instances, as well as when distinguishing between Neutral and Negative instances. This is evident from a relatively high number of misclassifications within these categories. These errors can likely be attributed to the restricted size of the training data available for these classes,

Class	%	Examples
Positive	21.3	• اي والله امان ياملكنا سلمان اعز الله بك الدين والاوطان • الهيئة طوق نجاة للأمة من تمسك بها نجى وأفلح • ابشروا بالخير
Negative	60.1	• ابلشونا بسماجهة مواضيعهم ونقاشاتهم الي مالها داعي • اتخيلني اسوق فالرياض بايعة عمري • ردو لي فلوسي حسبي الله من جفره لدحديره
Neutral	18.6	• تنويع مصادر الدخل امر ضروري • حراك وطني اقتصادي اجتماعي ثقافي هو انعكاس لبرنامج يتحدث عن المستقبل وفق رؤية استراتيجية تتسم بالشمولية العامه • ما قال شي غلط

Table 6: Statistics and example of tweets from the SDTC dataset (Al-Thubaity et al., 2018a).

as well as the difficulties faced by annotators during the process of labeling SDTC. Augmenting the training dataset with extra examples from both the Positive and Neutral classes could potentially enhance the accuracy of predicting these categories more precisely.

D Examples of Generated Tweets

Table 7 illustrates one example from each class generated by each language model for zero-shot scenarios discussed in Exp.3 in subsection 4.3.

Table 8 showcases examples where the two annotators answered each question with “No”.

Model	Examples
Bard AI	سعادة لا توصف بالحصول على وظيفة أحلامي Unbelievably happy to have my dream job مليت من الازدحام في شوارع الرياض Tired of crowding in the streets of Riyadh خلصت شغلي اليوم، الحين بروح أريح I finished my work today, now I'm going to rest
GPT-3.5	شاركت في فعالية تطوعية اليوم وشعرت بفخر كبير لخدمة مجتمعي. I participated in a volunteering event today and felt very proud to serve my community. تعبت من هالروتين اليومي البائس، لا متعة ولا فرحة في شيء! Tired of the miserable daily routine, there is no fun or joy in anything! الشاي الدافئ في الصباح يجعل يومي أفضل. Warm tea in the morning makes my day better.
GPT-4	أعشق التسوق في سوق الرياض، الباعة ودودون جداً وجودة البضائع عالية. I love shopping in the Riyadh market; the sellers are very friendly, and the quality of the goods is high. كل يوم أفقد الامل أكثر ولكن أصبر وأصبر ولا أرى طريق إلى السعادة. Every day, I lose hope more, but I am patient and patient, and I do not see a path to happiness. الحكومة السعودية تعمل بجد لتحسين البيئة التعليمية وذلك من خلال تحديث المناهج وتقديم التقنيات الحديثة. The Saudi government is working hard to improve the educational environment by updating curricula and introducing modern technologies.

Table 7: Examples from each class generated by each language model for zero-shot scenarios. The order of examples: positive, negative, and neutral.

Model	Questions	Example
Bard AI	Q1	أكل الكبسة السعودية من ألذ الأكلات في العالم. Eating Saudi Kabsa is one of the most delicious foods in the world.
	Q2	قررت أن أبدأ مشروعاً جديداً. I decided to start a new project.
	Q3	أحس أني عايش في حلم، ما أدري وش الواقع. I feel like I'm living in a dream, I don't know what the reality is. *Bard classified it as negative, while annotators classified it as neutral.*
	Q1+Q2+Q3	السياسة صعبة كثير. Politics is very difficult.
GPT-3.5	Q1	السفرة كانت كارثة، الأكل طعمه مرة يرفع الراس. تجربة سيئة جداً. The trip was a disaster, the food tasted bitter and raised the head. Very bad experience.
	Q2	قاعد برواق ع الهاش تقلب علي فرح وتسألني لية صفا مع انو جسدك حجر وتدور طول الليل تلف وعليسا عين جنوبية نوفة للفارس الكبير So-and-so, you are a loser and a loser next to you. You cannot do anything right in your life.
	Q3	المشتريات في هذا المول مكلفة جداً Purchases in this mall are very expensive *GPT-3.5 classified it as neutral, while annotators classified it as negative.*
	Q1+Q2+Q3	ويبدو ان الخطأ الاداري هذا تسبب في درجة حرارة البلاد بسنوات تسببت بيها حرارة اخطار المملكة العربية السعودية و الله قلل من نار It seems that this administrative error caused the country's temperature in the streets for years, caused by the heat of the dangers of the Kingdom of Saudi Arabia, and God reduced the fire of stone
GPT-4	Q1	في الحقيقة، الحده ليله ممتعة ببعض الأحيان، تجلس وحدك وما عندك أي مشاكل و أي شي يشغل بالك... حالة من الراحة و التفكير العميق. #راحة In fact, the intensity is a fun night sometimes, you sit alone and you don't have any problems or anything to worry about... a state of rest and deep thought. #comfort
	Q2	لديها التعليقات الجديدة على فيديو يوتيوب للشيخ سليمان الدويش، تحدث فيه عن المسائل الدينية اليومية. She has the new comments on a YouTube video by Sheikh Suleiman Ad-Dawish, in which he talks about everyday religious issues.
	Q3	بفضل تكنولوجيا اليوم، يمكننا التواصل مع الأشخاص في جميع أنحاء العالم بفضل لمسة واحدة على شاشات الهواتف الذكية. Thanks to today's technology, we can communicate with people all over the world thanks to a single touch on our smartphone screens. *GPT-4 classified it as neutral, while annotators classified it as positive.*
	Q1+Q2+Q3	زرت الرياض في نهاية الأسبوع واستمتعت في جولة في القصور والمعابد، كل شيء كان جميل ورائع. #السياحة_في_السعودية I visited Riyadh at the end of the week and enjoyed a tour of the palaces and temples, everything was beautiful and wonderful. #Tourism_in_Saudi Arabia

Table 8: Examples where the two annotators answered each question with "No".