mathematics

MDPI

Article

Switch-Transformer Sentiment Analysis Model for Arabic Dialects That Utilizes a Mixture of Experts Mechanism

Laith H. Baniata * and Sangwoo Kang *

School of Computing, Gachon University, Seongnam 13120, Republic of Korea
* Correspondence: laith@gachon.ac.kr (L.H.B.); swkang@gachon.ac.kr (S.K.)

Abstract: In recent years, models such as the transformer have demonstrated impressive capabilities in the realm of natural language processing. However, these models are known for their complexity and the substantial training they require. Furthermore, the self-attention mechanism within the transformer, designed to capture semantic relationships among words in sequences, faces challenges when dealing with short sequences. This limitation hinders its effectiveness in five-polarity Arabic sentiment analysis (SA) tasks. The switch-transformer model has surfaced as a potential substitute. Nevertheless, when employing one-task learning for their training, these models frequently face challenges in presenting exceptional performances and encounter issues when producing resilient latent feature representations, particularly in the context of small-size datasets. This challenge is particularly prominent in the case of the Arabic dialect, which is recognized as a low-resource language. In response to these constraints, this research introduces a novel method for the sentiment analysis of Arabic text. This approach leverages multi-task learning (MTL) in combination with the switch-transformer shared encoder to enhance model adaptability and refine sentence representations. By integrating a mixture of experts (MoE) technique that breaks down the problem into smaller, more manageable sub-problems, the model becomes skilled in managing extended sequences and intricate input–output relationships, thereby benefiting both five-point and three-polarity Arabic sentiment analysis tasks. The proposed model effectively identifies sentiment in Arabic dialect sentences. The empirical results underscore its exceptional performance, with accuracy rates reaching 84.02% for the HARD dataset, 67.89% for the BRAD dataset, and 83.91% for the LABR dataset, as demonstrated by the evaluations conducted on these datasets.

Keywords: switch transformer; mixture of experts (MoE) mechanism; sentiment analysis (SA); Arabic dialects; five-polarity; MTL

MSC: 68T07

Citation: Baniata, L.H.; Kang, S. Switch-Transformer Sentiment Analysis Model for Arabic Dialects That Utilizes a Mixture of Experts Mechanism. *Mathematics* **2024**, *12*, 242. https://doi.org/10.3390/math12020242

Academic Editors: Nebojsa Bacanin and Catalin Stoean

Received: 31 October 2023
Revised: 28 December 2023
Accepted: 8 January 2024
Published: 11 January 2024

1. Introduction

Sentiment analysis includes the computational process of discerning and understanding the emotional undertones or sentiments conveyed within a text, whether they be in the form of sentences, documents, or social media posts. This procedure aids businesses in acquiring insights into how their brands, products, and services are perceived, achieved through the evaluation of feedback from online interactions with customers. Platforms, such as Twitter, experience a significant daily influx of user-generated content in Arabic and Arabic dialects, and this trend is anticipated to endure as user-generated content continues its upward trajectory in the years to come. Opinions articulated in the Arabic language are estimated to account for approximately five percent of the linguistic landscape on the Internet. Additionally, Arabic has become one of the most influential languages online in recent times. It serves as a global language spoken by over five-hundreds million individuals worldwide and is categorized within the semantic language group. The official language across more than 21 countries ranges from the Arabian Gulf to the Atlantic Ocean,

and it is Arabic. Linguistically, Arabic stands out for its intricate complexity, setting it apart from English, largely due to its diverse range of dialects. The notable differences between Modern Standard Arabic (MSA) and Arabic dialects (ADs) add an extra layer of complexity. Furthermore, in the realm of Arabic language usage, the phenomenon of diglossia is widespread. This means that in in informal settings, individuals use local Arabic vernaculars, while in formal or professional settings, Modern Standard Arabic (MSA) is employed. For instance, depending on the circumstances, individuals in Libya may switch between MSA and their native Libyan dialects. The Libyan dialect encapsulates the nation's historical narrative, cultural identity, heritage, and shared life experiences. Diverse regional expressions in Arabic exhibit noticeable disparities in their geographical distribution, including Levantine (encompassing Palestine, Jordan, Syria, and Lebanon), Maghrebi (covering Morocco, Algeria, Libya, and Tunisia), Iraqi, Nile Basin variants (found in Egypt and Sudan), and Arabian Gulf versions (extending across the UAE, Saudi Arabia, Qatar, Kuwait, Yemen, Bahrain, and Oman). Discerning emotionally charged terms amidst this wide array of Arabic linguistic diversity presents a considerable obstacle due to the language's intricate structural attributes, orthography, and overall intricacy. Each nation where Arabic is spoken showcases its own unique colloquial language, further augmenting the intricateness of the linguistic landscape. To illustrate, Arabic content disseminated on social media platforms frequently merges Modern Standard Arabic (MSA) with regional dialectal Arabic, resulting in distinct interpretations of the same word.

Moreover, an additional syntactic challenge in Arabic dialects (ADs) pertains to word arrangement. In order to analyze this issue, it is crucial to understand the arrangement of the verb, subject, and object in an AD sentence. As highlighted in the literature review, languages can be categorized into different groups, such as subject–object–verb as seen in Korean, subject–verb–object as seen in English, and verb–object–subject as is the case with Arabic [1]. Additionally, there are languages that allow for a flexible word order such as ADs [2]. Within AD expressions, this flexibility imparts advanced insights into the subject, object, and various other forms of information. Hence, the utilization of a single-task learning methodology and depending exclusively on manually created features are insufficient for carrying out sentiment analyses on Arabic dialects. Furthermore, these divergences within Arabic dialects (ADs) present a significant challenge for standard deep learning algorithms. This is due to the fact that longer phrases in ADs introduce a wealth of complex and confusing contextual information related to the object, verb, and subject. An issue with traditional deep learning methods is the depletion of input sequence data, resulting in the reduced effectiveness of the sentiment analysis (SA) model as the length of the input sequence increases. Also, the configuration of Arabic words' roots and characters can vary significantly depending on the context, as exemplified by (Ketab كتاب, Ketabat كتابات, Yaktob يكتب). Moreover, the absence of standardized orthographic conventions stands out as a primary challenge in ADs. This encompasses morphological distinctions across dialects, evident in the utilization of prefixes and suffixes absent from Modern Standard Arabic (MSA). Additionally, it is important to highlight that many Arabic words can express various meanings, depending on the use of diacritics within the same syntactical structure. Additionally, the development of deep learning-powered sentiment analysis (SA) models necessitates a substantial corpus of training data, a resource that proves challenging to amass for Arabic dialects (ADs). These dialects are recognized as unstructured and resource-scarce languages, rendering the retrieval of information a formidable endeavor [3]. As the quantity of training data decreases for Arabic dialects (ADs), the effectiveness of classification also diminishes. Furthermore, most tools designed for Modern Standard Arabic (MSA) fail to consider the unique characteristics of Arabic dialects [4]. It is also important to note that relying solely on lexical resources, like lexicons, may not be the most effective approach for Arabic SA due to the vast array of words stemming from diverse dialects, making it improbable for any lexicon to encompass them all [5]. Furthermore, the creation of tools and resources tailored to Arabic dialects is a laborious and time-intensive undertaking [6].

Lately, there has been a heightened emphasis on exploring sentiment analysis in the Arabic language. The research primarily centers on classifying opinions and tweets to detect both binary and ternary emotional tones. Most of these approaches [7–12] depend on lexicons and attributes specific to tweets, which function as inputs for machine learning (ML) algorithms. In contrast, alternative methods embrace a rule-based approach, such as employing principles of lexicalization. This involves establishing and prioritizing a set of heuristic rules to effectively categorize tweets into negative or positive sentiments [13]. Shifting focus, the Arabic sentiment ontology introduces sentiments with diverse levels of intensity to distinguish user attitudes and streamline the classification of tweets. Deep learning approaches for sentiment analysis, encompassing RNNs [14], CNNs [15–18], and recursive auto-encoders, have attracted substantial interest because of their impressive flexibility and robustness achieved via automated feature extraction. Notably, the recently developed switch-transformer model [19] outperforms conventional transformer models [20], recurrent neural network (RNN)-based models in various natural language processing (NLP) tasks, thereby capturing the interest of researchers in the field of deep learning.

This research paper proposes a switch-transformer sentiment analysis (ST-SA) model, that utilize an MoE mechanism that breaks down the problem into smaller, more manageable sub-problems. In every expert layer, the router decides which expert will receive the token. This choice is made from among the available experts, depending on the characteristics of the token's representation. The router selects the most suitable expert based on the current representation of the incoming token. However, it lacks counterfactual data regarding the potential effectiveness of choosing a different expert. The proposed model becomes adept at handling lengthy sequences and intricate input–output relationships, benefiting both five-point and three-polarity Arabic sentiment analysis tasks. Despite previous efforts to address the challenges of AD SAs, the approach of MTL has emerged as a promising solution.

Multi-Task Learning

Multi-task learning (MTL) in deep learning models is a powerful approach that aims to improve learning efficiency and prediction accuracy by simultaneously training a single model on multiple related tasks. This technique leverages the commonalities and differences across tasks to enable a model to generalize better for each task. In MTL, tasks share representations, allowing the model to exploit the useful information presented in related tasks, thereby reducing the risk of overfitting on any single task. This is particularly beneficial when the data for some tasks are scarce. MTL models often use shared layers for learning common features while employing task-specific layers to capture the unique aspects of each task. This shared learning leads to more robust models that are capable of handling a variety of challenges. As a result, multi-task learning has found applications in numerous fields, such as natural language processing, where a single model can simultaneously learn tasks, like sentiment analysis, language translation, and named entity recognition, leveraging the synergies between these tasks to enhance the performance. MTL enriches comprehension capabilities, elevates the encoder quality, and augments the significance of sentiment classification compared to a conventional single-task classifier. This is accomplished by simultaneously handling interconnected tasks and utilizing a common representation of text sequences [21]. An essential benefit of multi-task learning (MTL) lies in its capability to efficiently utilize diverse resources for similar tasks. However, it is noteworthy that most existing approaches for SAs of ADs predominantly focus on binary and ternary classifications. In this study, we redirect our attention to the five-polarity AD SAs problem, an area that, to our knowledge, has received limited attention. Notably, the utilization of a switch-transformer architecture in conjunction with MTL for AD SA classifications has not been explored in prior studies. Previous methodologies addressing this classification primarily relied on a conventional transformer and Bi-LSTM techniques. We can summarize our contributions as follows:

- This research article introduces a pioneering switch-transformer model that uses MTL for SAs of ADs. The proposed ST-SA model founded on the MoE mechanism is developed to break down the problem into smaller, more straightforward sub-problems, enabling the model to effectively handle extended sequences and intricate input–output connections.
- Furthermore, a multi-head attention (MHA) mechanism is devised to capitalize on the correlation between three and five polarities through the utilization of a shared switch-transformer encoder layer. We clarify the method of sequentially and collaboratively mastering two tasks (ternary and five classifications) within the multi-task learning (MTL) framework. This strategy is designed to enhance the representation of Arabic dialect (AD) texts for each task and broaden the range of captured features.
- This research paper studies the effect of training the proposed switch-transformer model with varying embedding dimensions for each token, diverse token values, different attention head numbers, varying filter sizes, a diverse number of experts, a range of batch sizes, and multiple dropout values.
- The proposed SA-ST model employs an MHA mechanism to evaluate the correlation strength between two words within a sentence. This notably bolsters the relevance and importance of various natural language processing tasks.

The following sections of this paper are organized in the following manner: Section 2 offers a literature overview, Section 3 provides a comprehensive explanation of the proposed model, Section 4 presents the experimental results, and, finally, Section 5 summarizes the conclusions derived from this study.

2. Literature Review

The research focusing on five levels of polarity classification tasks in Arabic sentiment analyses has received relatively less attention compared to binary or ternary tasks. Additionally, the majority of approaches addressing this particular task rely on traditional machine learning algorithms. For example, methods utilizing corpora and lexicons were examined by incorporating bag of words (BoW) features along with various machine learning algorithms, such as passive aggressive algorithm (PA), support vector machine (SVM), logistic regression (LR), naive Bayes (NB), perceptron, and stochastic gradient descent (SGD) for analyses regarding Arabic book reviews [22]. Similarly, [23] explored the impact of stemming and the balancing of BoW features using multiple machine learning algorithms on the same dataset. They found that applying stemming resulted in a decline in performance. In [24], a divide-and-conquer approach was proposed to handle tasks related to the ordinal-scale classification. Their model adopted a hierarchical classifier (HC) structure, breaking down the five labels into smaller sub-problems. It was noted that the HC model surpassed a single classifier. Expanding on this foundation, various architectures for hierarchical classifiers were introduced [25]. These structures were compared against machine learning classifiers, such as SVM, KNN, NB, and DT. The experimental results indicate an improvement in performance with the hierarchical classifier. Nevertheless, it is important to mention that many of these structures exhibited a reduction in performance.

In a different study [26], an examination focused on diverse machine learning classifiers, including LR, SVM, and PA, utilizing n-gram attributes within the context of book reviews in the Arabic dataset (BRAD). The results showed that SVM and LR presented the most commendable performances. Similarly, [27] conducted an assessment on multiple sentiment classifiers, encompassing AdaBoost, SVM, PA, random forest, and LR, using the hotel Arabic reviews dataset (HARD). Their observations revealed that SVM and LR exhibited superior performances, particularly when incorporating n-gram features. These mentioned approaches underscore a significant lack of deep learning strategies for the classification of five polarities in Arabic sentiment analysis (SA). Additionally, a majority of the methods dealing with these five polarity tasks are rooted in traditional ML algorithms, relying on the feature engineering process, known for its time-consuming and challenging nature. Furthermore, these approaches are built upon single-task learning (STL) and lack

the capability to discern the interrelationship among different tasks (cross-task transfer) and model various polarities concurrently, including both five and three polarities.

Other investigations have turned to MTL to tackle the challenge of five-point SA classification tasks. For instance, [28] introduced a multi-task learning framework utilizing a recurrent neural network (RNN) to concurrently address both five-point and ternary classification tasks. Their model incorporated bidirectional long short-term memory (Bi-LSTM) and multilayer perceptron (MLP) layers. Additionally, they enriched features with tweet-specific elements, like punctuation counts, elongated words, emoticons, and sentiment lexicons. Their findings indicate that jointly training SA classification tasks significantly boosts the efficacy of the five-polarity task. Similarly, in [29], the synergy between five-polarity and binary sentiment classification tasks was harnessed through concurrent training. The proposed model incorporated an encoder (LSTM) and a decoder (variational auto-encoder) as shared components for both tasks. The empirical results highlight that the multi-task learning (MTL) model improved the performance for the five-polarity task. The concept of adversarial multi-task learning (AMTL) was first introduced in [21]. This model integrates two LSTM layers as task-specific components and one shared LSTM layer across tasks. Additionally, a convolutional neural network (CNN) was fused with the LSTM, and the outputs from both networks were concatenated with the shared-layer output, forming the final latent sentence representation. The authors observed that the proposed multi-task learning (MTL) model enhanced the performance of five-polarity classification tasks and enhanced the quality of the encoder. While the multi-task learning (MTL) approaches detailed above have found application in English, there is a noticeable lack of multi-task learning and deep learning techniques applied to five-polarity Arabic sentiment analysis (SA). Existing studies concentrating on this task predominantly rely on single-task learning with traditional machine learning algorithms. Consequently, there is ample room to enhance the effectiveness of current Arabic SA methods in addressing the five polarities, as it remains at a relatively modest level.

Subsequent inquiries have utilized advanced deep learning methodologies for the analysis of sentiments (SAs) in various fields, including finance [30,31], movie critiques [32–34], weather-related tweets [35], reviews on travel platforms [36], and cloud service recommendation systems [37]. Numerous studies have harnessed polarity-based sentiment deep learning techniques for analyzing tweets [38,39]. A multitude of techniques have been proposed for emotion recognition [40,41]. In the realm of dialogue emotion recognition, Wang [42] introduced the hierarchically stacked graph convolution framework. This framework aims to improve the extraction of discriminative information from the emotional graph it constructs. To achieve this, it incorporates the potent transformer operation along with a residual connection. The efficacy of this method was substantiated through comparative experiments conducted on the IEMOCAP dataset.

Wang [43] developed "NUAN", a non-uniform attention network, to integrate multimodal features effectively. NUAN utilizes an attention mechanism that focuses differently on three types of data: text, which is treated as a fixed representation, and acoustic and visual data, which are used to enrich the text-based information in a structure called the tripartite interaction representation. This network incorporates a unique non-uniform attention module within the LSTM (long short-term memory) framework, allowing it to process data over successive time steps. The LSTM and the non-uniform attention module's (NUAM's) outputs are merged into a single vector and fed through a linear embedding layer to perform the final sentiment analysis. The method's effectiveness is validated through tests on two different databases. Baniata et al. [44] introduced a novel approach utilizing a multi-task learning multi-head attention model for the five-point classification of ADs. This innovative architecture incorporates a self-attention technique and a multi-task learning (MTL) framework to bolster the overall representation of text sequences. Moreover, the self-attention method enables the selection of the most pertinent terms and phrases from these sequences. By training for sentiment analysis (SA) tasks, encompassing ternary and five-polarity tasks specific to ADs, the system's efficacy was significantly elevated.

Leveraging the advantages of self-attention and MTL amplified the proficiency of the proposed SA system. The outcomes of this study underscore the pivotal attributes of the MTL self-attention SA system, which leverages the self-attention method and increases the accuracy of the results for both five-point and three-point classification tasks. The incorporation of the MTL framework and word units as input features for the self-attention sub-layer indicates their critical role in low-resource language SA tasks, such as those involving Ads. Additionally, refining the model through diverse configurations, including employing multiple heads in the self-attention sub-layer and training with multiple encoders, notably enhanced the classification performance of the suggested system. Furthermore, Alali et al. [45] introduced a multi-tasking methodology termed the multi-task learning hierarchical attention network (MTLHAN). This approach aims to augment the representation of sentences and enhance the overall adaptability. The MTLHAN framework employs a shared word encoder and attention network for both tasks, utilizing two different training strategies to scrutinize three-polarity and five-polarity Arabic sentiments. The outcomes of the experiments emphasize the outstanding performance of this suggested model.

Singh et al. [46] set out to compare unsupervised lexicon-based models (Text Blob, AFINN, and Vader Sentiment) with supervised machine learning models (KNN, SVM, random forest, and naive Bayes) in their research paper focused on sentiment analysis. Notably, this study marks the first investigation concentrating on cyber public opinion related to the abrogation of Article 370. The researchers amassed Twitter data, comprising over 200,000 tweets, from which 29,732 tweets underwent selection for analysis post-data cleaning. The findings reveal that, among the supervised learning models, random forest exhibits the most exceptional performance. Conversely, within the unsupervised learning models, Text Blob attained the highest accuracy, registering 99% and 88% accuracy values, respectively. It is noteworthy that the performance metrics of the proposed supervised machine learning models surpass the outcomes of a recent sentiment analysis study conducted in 2023. Table 1 presents an overview of the sentiment analysis methods employed for Arabic dialects.

Table 1. Sentiment analysis methods employed for Arabic dialects.

Technique	Model	Dataset (5 Polarity)	Ref.
Corpora and Lexicons	SVM, LR, NB, PA	LABR	[22]
BoW	SVM, LR, NB, KNN, J48, C4.5, DT	LABR	[23]
Divide and Conquer	Hierarchical Classifier (HC)	LABR	[24]
N-gram	LR, SVM, PA	BRAD	[26]
N-gram	AdaBoost, SVM, PA, RF, LR	HARD	[27]
Multi-Task Learning	Transformer	HARD, BRAD, LABR	[44]
Multi-Task Learning	Hierarchical Attention over Bi-LSTM	HARD, BRAD, LABR	[45]

3. The Proposed Switch-Transformer Sentiment Analysis Model That Utilizes the MoE Mechanism

Transformer-based models have exhibited remarkable efficacy across a spectrum of NLP tasks, encompassing the categorization of text. The conventional transformer architecture [20], featuring multi-head attention, is a prevalent blueprint for this endeavor. As illustrated in Figure 1, its composition includes an encoder composed of multiple layers of multi-head attention (MHA) and feedforward neural (FFN) networks. This multi-head attention (MHA) method grants the model the capacity to assess the significance of various terms in a sequence grounded on their semantic associations, while the FFNs convert the output of the MHA layer into a more advantageous representation. The crux of the transformer is the MHA method founded on mathematical expressions [47]. Presented with a succession of input embedding values, $x_1, \ldots, x_n$, the MHA method derives a collection of contextually attuned embeddings, $h_1, \ldots, h_n$, through the ensuing procedure:

$$h_i = Attention\left(QW_i^Q, KW_i^k, VW_i^V\right) \quad (1)$$

where attention is the scaled dot-product attention function:

$$Attention\ (Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Subsequently, the multi-head attention (MHA) consists of the concatenation of all its heads, h_i, as follows:

$$Multihead\ (Q, K, V) = concat(h_1, \ldots\ldots, h_n)\ W^o \tag{3}$$

Moreover, the position-wise feedforward networks (FFNs) refer to multi-layer perceptron that operate individually on each position within the sequence. These FFNs deliver a non-linear transformation of the attention outputs, and their calculation follows this pattern:

$$FFN(x) = ReLU(xW_1 + b_1)\ W_2 + b_2 \tag{4}$$

In every layer, a layer normalization is applied to normalize the inputs within a neural network, enhancing the speed and stability of the training.

$$LayerNorm\ (x) = \gamma\frac{x - \mu}{\sqrt{\sigma^2 + \in}} + \beta \tag{5}$$

In this context, Q, K, and V represent the query, key, and value matrices, respectively, while W_i^Q, W_i^k, and W_i^v signify the weight matrices that have been acquired through learning for the specific head denoted as i within the multi-head attention mechanism. W_1 and W_2 are the weight matrices pertaining to the position-wise feedforward networks (FFNs), and γ and β denote the acquired scaling and shifting parameters used for the layer normalization. Additionally, μ and σ refer to the mean and standard deviation, respectively, of the feature activations in the input. The operational process within the transformer structure can be succinctly summarized through the ensuing steps:

- Linear transformation: the sequence of the input undergoes a transformation, resulting in the creation of three vectors: query Q, key K, and value V. This is achieved through the application of a linear transformation to the embedded input.
- Segmentation: the vectors Q, K, and V are subsequently divided into multiple heads denoted as h_i. This enables the model to concurrently attend on distinct facets of the input sequence, as described in Equation (1).
- Scaled dot-product attention: for every h_i, the model determines the attention weights between the Q and K vectors by proportionally adjusting their dot products using the square root of the vector dimension. This process evaluates the significance of each K vector in relation to its corresponding Q vector.
- SoftMax: the resultant attention weights undergo normalization through the application of a SoftMax function, guaranteeing that their collective sum amounts to 1.
- The attention weights are subsequently employed to balance the V vectors, generating an attention output for each component, h_i, as indicated in Equation (2).
- The combined attention outputs from each head are merged and then re-mapped to the initial vector dimension via an additional linear transformation, as outlined in Equation (3).
- Feedforward network: the resulting outcome undergoes a transmission through a forward-propagating network, introducing nonlinearity and enabling the model to identify more intricate connections between the input and output, as stated in Equation (4).

By applying these procedures to every layer within both the encoder, the MHA mechanism empowers the transformer framework to identify intricate semantic connections among words in a sequence, proving highly efficient across various natural language processing tasks. Nevertheless, the conventional transformer design encounters specific limitations. A primary concern revolves around the MHA mechanism's quadratic computational demand concerning the input sequence length, hindering the scalability for exceptionally long sequences [48] and decreasing the adaptability for shorter sequences. Furthermore, the MHA treats all positions in the input sequence uniformly, which might not be optimal for specific input types where certain positions hold greater significance than others. While the transformer system demonstrates an outstanding performance for numerous NLP tasks, it may still struggle with capturing intricate input–output associations that necessitate the use of more specialized models.

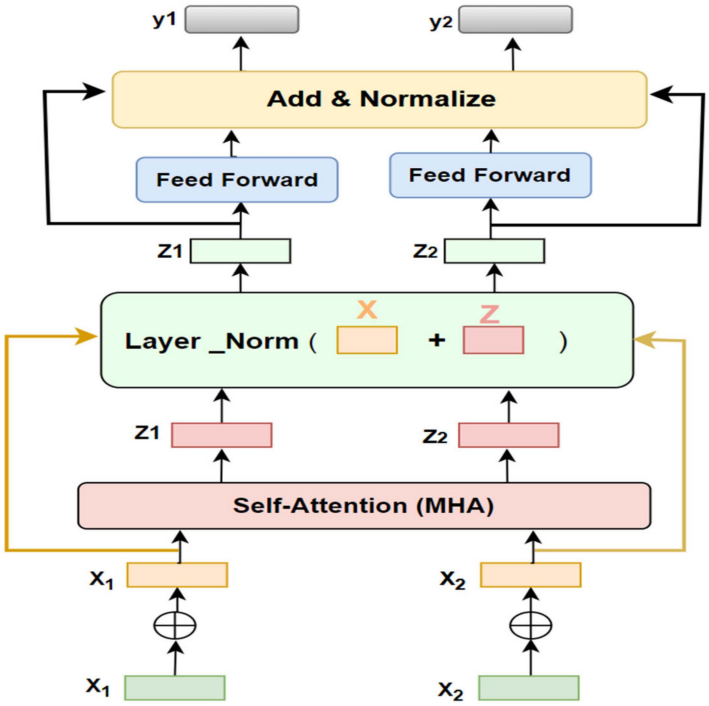

Figure 1. Architecture of the conventional transformer model.

To surmount these obstacles, our research paper introduced a novel switch-transformer sentiment analysis (ST-SA) model, employing multi-task learning (MTL) for the classification of Arabic dialect (AD) sentiments. The objective behind adopting multi-task learning (MTL) was to augment the performance of the five-point Arabic sentiment analysis quandary, capitalizing on the interconnection between the AD SA classification tasks, encompassing both five and ternary polarities. The proposed ST-SA model for ADs was based on the transformer model recently elucidated by Vaswani et al. [20]. MTL exhibits greater efficacy compared to singular-task learning. It harnesses the communal representation of diverse loss functions, concurrently handling SA tasks with three and five polarities, thereby refining the representation of both the semantic and syntactic facets of AD texts. The insights garnered from each task can fortify the learning process of other tasks, enhancing their efficacy. Furthermore, a pivotal facet of MTL lies in its a superior approach to accessing resources devised for akin tasks, ultimately amplifying the learning proficiency of the current task and enriching the reservoir of exploitable knowledge. By means of comprehension, the layers involved in task sharing can amplify the model's capacity for generalization, accelerate the pace of learning, and enhance its overall intelligibility. Similarly, leveraging the domain expertise embedded in the training cues of interconnected tasks as an inductive bias, the multi-task learning approach facilitates swift transfers that

bolster generalization. This inductive transfer can be deployed to refine the precision of generalization, expedite the learning process, and heighten the transparency of the acquired models. A learner engaged in the simultaneous acquisition of numerous interrelated tasks can employ these tasks as an inductive bias for one another, thereby gaining a more profound understanding of the domain's regularities. This can result in a more practical acquisition of sentiment analysis (SA) tasks for Arabic dialects (ADs) even with a limited amount of training data. Similarly, multi-task learning collaboratively discerns the meaningful interrelation among the acquired tasks. As depicted in Figure 2, the proposed ST-SA sentiment analysis system boasts a distinctive architecture relying on multi-head attention (MHA), MTL, a shared vocabulary, and specialized mechanisms referred to as a mixture of experts (MoE) inside the switching FNN layer. The presented ST-SA model, employing MTL, fine-tunes mixed classification tasks (ternary and five-polarity classification tasks) and comprehends them collectively. The integration of a shared switch-transformer block (encoding layer) streamlines the transfer of knowledge from the ternary task to the five-point task during the learning process, leading to an enhancement in the current task's (five-point task) learning capabilities.

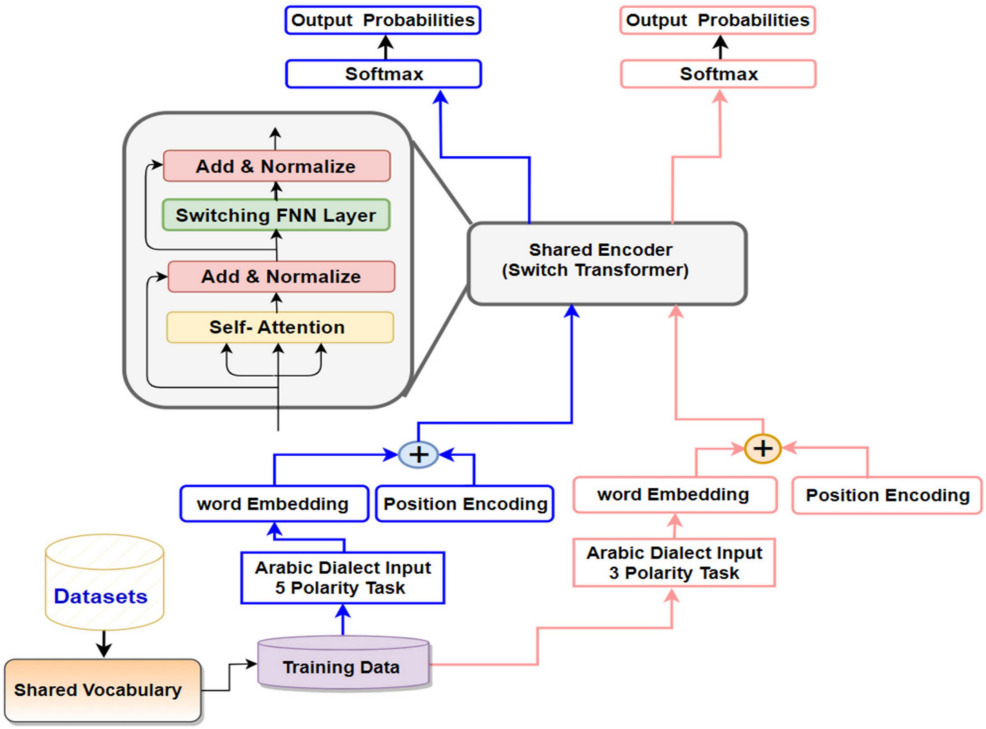

Figure 2. The Architecture of the proposed switch-transformer model for Arabic dialects that utilizes the MoE mechanism.

Switch transformers (STs) [19] endeavor to address the limitations found in traditional transformer architectures by integrating a specialized system known as the mixture of experts (MoE). This tactic decomposes complex problems into smaller, more digestible components, thus enhancing the model's capability to manage longer sequences and complex relationships between inputs and outputs. Notably, the multi-head self-attention mechanism in standard transformers, while adept at identifying semantic relationships in sequences, falters with shorter lengths. The MoE framework allows the model to divide sequences into smaller, more tractable portions, assigning specific expert networks to each, subsequently improving the efficiency and accuracy for tasks with shorter sequences, as evidenced by its superior performance in several benchmark tests [49–51]. A critical innovation in switch transformers compared to the original model is the replacement of the conventional feed-forward network (FFN) with the MoE mechanism, as depicted in Figure 3. Traditional transformers utilize an FFN consisting of two linear layers and a

ReLU activation function. However, the MoE approach utilizes a suite of expert networks that each analyze different aspects of the input data, integrating their results through a gating network. This allows the model to dynamically adapt and select from a variety of parameter sets or expert modules tailored to the specific input, a marked departure from the fixed parameter approach of the traditional transformer, as described in Equation (4). In a formal manner, the MoE mechanism within the switch transformer can be denoted by the subsequent equation:

$$z_t = \sum_j \ g_j(x_t) * e_j(x_t) \tag{6}$$

The function $g_j(x_t)$ serves as a gate, influencing the significance of expert module j with respect to input x_t. Meanwhile, $e_j(x_t)$ represents the result produced by expert module j for input x_t. The switch mechanism operates by training the gating functions' parameters, which enable the dynamic selection of expert modules. This adaptive capability equips the model to accommodate diverse input patterns and excel across a range of tasks. The main function of MoE mechanism within the switch transformer can be summarized in several steps. First, the input undergoes partitioning into various subspaces, with each subspace undergoing individual processing by a distinct expert. Each of these experts constitutes an independent neural network that is trained to excel in a particular subset of the input domain. Each expert generates an output vector that presents its forecast for the specific input subspace provided. A gating process is employed to identify the expert most pertinent to a given input. This gating process takes the input and generates a series of weights that ascertain the significance of each expert's prediction. The final output is a weighted combination of the experts' forecasts, and the weighting for this amalgamation is dictated by the gating mechanism, and the ST-SA model training cycle is summarized in Algorithm 1. MTL can be implemented by sequentially engaging the loss and optimizer for each task. This entails running the training for a predetermined number of cycles on the ternary classification task, then shifting focus to the five-polarity classification tasks. The objective of training both tasks is to minimize cross-entropy. Consequently, we achieve:

$$\hat{y}_{(ternary)} = \text{softmax}\left(W_{(ternary)} s^s_{it(ternary)} + b_{(ternary)}\right), \tag{7}$$

$$\hat{y}_{(five)} = \text{softmax}\left(W_{(five)} s^s_{it(five)} + b_{(five)}\right), \tag{8}$$

where $\hat{y}^i_j$ and y^j_i are the anticipated likelihoods and ground-truth labels, respectively. N_1 and N_2 are the numbers of training samples in five-point and ternary classification tasks, respectively. In order to implement the joint training of five-point and ternary classifications to train the ST-SA system, we received the following global loss function:

$$Total\ Loss(L) = \lambda_1 L_{ternary}(\hat{y}, y) + \lambda_2 L_{five}(\hat{y}, y), \tag{9}$$

where λ_1 and λ_2 are the weighs for the five-point and ternary classification tasks, respectively. Parameters λ_1 and λ_2 are utilized to balance both losses using the equal-weighting strategy ($\lambda = 1$). In general, the MoE empowers the switch transformer to master intricate patterns within the input domain by capitalizing on the specialized expertise of numerous experts. This framework enables the model to glean insights from multiple experts, each adept in distinct facets of the data, and fuse their results to enhance the overall performance. This can result in superior proficiency in tasks demanding a thorough comprehension of inputs, presenting a hopeful remedy for the constraints of limited datasets in AD text classifications. Consequently, the study leverages this capacity to discern intricate relationships among words and phrases within AD texts.

Algorithm 1: Switch Transformer-Sentiment Analysis Model for Arabic Dialects that Utilizes the MoE Mechanism

Require: training dataset $X\left(X_{(ternary)}, Y_{(ternary)}, X_{(five)}, Y_{(five)}\right)$ learning rate l;
Ensure: model Ω: $\left\{W_{(ternary)}, W_{(five)}, b\right\}$;
1: Initialize model Ω: $\left\{W_{(ternary)}, W_{(five)}, b\right\}$;
2: Iterate
3: Pick a ternary task
4: From Task K: Pick mini-batch samples
5: Ternary classification: $X_{(ternary)}$
    5.1: Self-attention sub-layer $(\alpha_{(ternary)})$
    5.2: Switching FFN sub-layer
6: Five-polarity classification: $X_{(five)}$
    6.1: Self-attention sub-layer $(\alpha_{(five)})$
    6.2: Switching FFN sub-layer
7: SoftMax_layer $(ternary)$
8: SoftMax_layer $(Five)$
9: If the training = jointly, then
10: Compute loss: $J(\Omega)$ using Equation (9)
11: else
12: if the training = alternately and task = ternary
13: Compute loss for every task: $J(\Omega)$ using Equation (7)
14: else
15: Compute loss for every task: $J(\Omega)$ using Equation (8)
16: Compute gradient: $\nabla(\Omega)$.
17: Update the model: $\Omega = \Omega - l\nabla(\Omega)$
18: Till reaching the max number of epochs

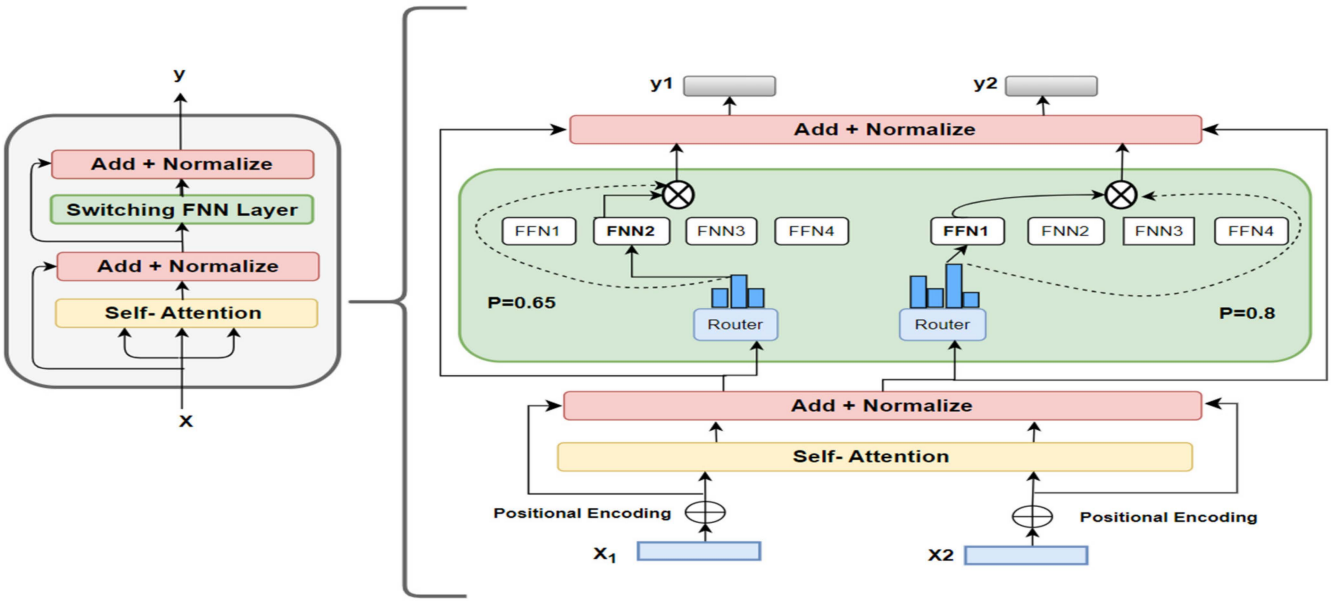

Figure 3. Detailed architecture of the encoder and switching FNN layer in the proposed ST-SA model.

4. Experiments

A series of practical tests were conducted to evaluate the effectiveness of the ST-SA model for Arabic dialect vernaculars. The performance of the proposed ST-SA model for classifying Arabic dialects (ADs) was thoroughly examined.

4.1. Data

The proposed model underwent training using three reference datasets. The initial dataset utilized was HARD [27], where reviews were gathered from multiple booking platforms and organized into five unique categories. The second dataset, BRAD [26], and the third dataset, LABR [22], were subsequently employed for training purposes. The research project utilized review-level datasets, including BRAD, HARD, and LABR. BRAD's reviews were collected from the Goodreads website and categorized into five scales. The distribution of classes for HARD, BRAD, and LABR is detailed in Tables 2–4 respectively. It is important to note that the datasets employed in this study were left in their unprocessed state, potentially affecting the reliability of the proposed model. Additionally, all sentences were processed through preprocessing steps, including the application of sentence segmentation tools to divide the reviews into individual sentences. This process also involved the complete removal of Latin alphabets, non-Arabic symbols, diacritical marks, hashtags, punctuation, and URLs from the AD texts. The texts in the Arabic dialects were subject to orthographic standardization to ensure uniformity [2]. Emoticons were translated into their text descriptions and modifications were made to elongated words. To avoid model overfitting, an early stopping strategy was employed with the patience parameter set for three epochs. In evaluating the performance of the developed ST-SA model, which employed MTL for the sentiment analysis of Arabic dialects, a model checkpoint system was used to store the most favorable weights of the model. The data were split, allocating 80% for training and 20% for testing, and a K-fold cross-validation was applied with k = 2 to create a training/testing split for evaluating the model [52]. The analyses of the HARD, BRAD, and LABR datasets provided insights into the distribution of sentiments across samples. The HARD dataset consisted of 409,562 entries divided into 5 categories of sentiments. By dedicating 80% of the data (327,649 samples) for training and the remaining 20% (81,912 samples) for testing, the model provided a thorough understanding of the range of sentiments. Similarly, the BRAD dataset, with 510,598 entries, followed the same distribution: 80% (408,478 samples) for training and 20% (101,019 samples) for testing. The LABR dataset, smaller, with 63,257 entries, also adhered to this 80–20 split for training (50,606 samples) and testing (12,651 samples). Such divisions ensured that the five sentiment categories were adequately represented in both the training and testing phases, enabling the models to learn the nuances of sentiment differences and effectively generalize this knowledge to new, unseen data. Prejudices can wield significant sway over the effectiveness of sentiment analysis models. If biases are present in the training data, they can skew the outcomes. To address this concern and determine the appropriate data selection for the presented ST-SA sentiment analysis model for Arabic vernaculars, we performed five distinct steps:

- Guaranteed that the training dataset comprised a multitude of origins and encompassed a broad spectrum of demographic profiles, geographic locales, and societal contexts. This approach served the purpose of mitigating biases, resulting in a dataset that was not only more exhaustive, but also more equitable in its composition.
- Confirmed that the sentiment labels in the training dataset were evenly distributed among all demographic segments and viewpoints.
- Set forth precise labeling directives that explicitly guided human annotators to remain impartial and refrain from introducing their personal biases into the sentiment labels. This approach aided in upholding uniformity and reducing the potential for biases.
- Conducted an exhaustive examination of the training data to pinpoint potential biases was imperative. This entailed scrutinizing factors, like demographic disparities, serotype reinforcement, and any groups that could be inadequately represented. Upon identification, we implemented appropriate measures to rectify these biases. This involved employing techniques, such as data augmentation, oversampling of underrepresented groups, and applying preprocessing methods.

Table 2. Statistics for HARD dataset.

Task Type	Highly Positive	Positive	Neutral	Negative	Highly Negative	Total
3 Polarity	-	132,208	80,326	38,467	-	251,001
5 Polarity	144,179	132,208	80,326	38,467	14,382	409,562

Table 3. Statistics for BRAD dataset.

Task Type	Highly Positive	Positive	Neutral	Negative	Highly Negative	Total
3 Polarity	-	158,461	106,785	47,133	-	251,001
5 Polarity	16,972	158,461	106,785	47,133	31,247	510,598

Table 4. Statistics for LABR imbalanced dataset.

Task Type	Highly Positive	Positive	Neutral	Negative	Highly Negative	Total
3 Polarity	-	15,216	9841	4197	-	29,254
5 Polarity	19,015	15,216	9814	4197	2337	50,606

4.2. The Setup of the Proposed Model

The introduced sentiment analysis model, known as the switch transformer that utilized multi-task learning (ST-SA), was created by harnessing the capabilities of TensorFlow [53], Keras [54], and scikit-learn [55] frameworks. To explore its effectiveness, a series of experiments were carried out for all ADs classification tasks, encompassing both three and five polarities. These experiments involved a diverse array of parameter configurations, specifically considering 6 different values for the word-embedding dimension of each token: 50, 32, 40, and 35. Additionally, the attention heads were assessed with 6 distinct values: 4, 2, and 3. The position-wise FNN incorporated filters of varying dimensions, including 40, 30, 35, 32, and 50.

4.3. The Training Mechanisim of the Proposed ST-SA Model for Arabic Dialects

Joint training and alternative training are two key approaches in the realm of multi-task learning models. Joint training involves training a model on multiple tasks simultaneously, sharing information and learning representations that are beneficial for all tasks. This approach leverages the interdependencies between tasks to improve the overall performance. In contrast, alternative training focuses on training the model on tasks individually, cycling through them iteratively. The suggested model is thus enabled to concentrate its attention on each task individually, which can result in an enhanced performance for separate tasks. Each methodology comes with its own set of benefits and disadvantages. Joint training can foster improved generalizations over multiple tasks, whereas alternating training can be more beneficial for tasks that greatly differ in their data distribution or complexity. The decision on which strategy to employ depends largely on the unique attributes of the tasks involved and the preferred balance between performance efficiency. In the end, choosing the right training approach is pivotal for determining the success and flexibility of multi-task learning models.

The designed system efficiently managed both ternary and five-level sentiment classification tasks. For example, when training with the HARD dataset, ST-SA toggled between teaching the model to understand both the five-level and ternary classification tasks. We explored two training methodologies: an alternating [2, 1] method and a simultaneous joint learning approach. In our approach to multi-task learning, we sequentially applied the loss function and optimizer to each task. This approach involved beginning the training with the ternary classification task for a set number of epochs and then shifting the focus to the five-polarity task, with an overarching goal to reduce categorical cross-entropy across both tasks. The ST-SA model was trained over 20 epochs, integrating an early stopping feature

set to trigger after two epochs of no improvement, and the batch size was set to 90. We adhered to the established guidelines for the BRAD, HARD, and LABR datasets, dividing them into 80% for training and 20% for evaluation. The Adam optimizer was selected for guiding each task within the ST-SA framework. Sentence segmentation was employed to divide reviews into individual sentences, adhering to maximum sentence-length standards specific to each dataset: 80 for BRAD, 50 for HARD, and 80 for LABR. We did not apply class weights to our model [56]. To ensure robust learning, the training data were shuffled before each epoch. Further specifics regarding the hyper-parameters and their settings are detailed in Section 4.5.

4.4. Sate-of-the-Art Approaches

Employing the five-point datasets, BRAD, HARD, and LABR, for analyzing ADs, the ST-SA model designed for this purpose was assessed against the latest standard methods. Originally, the use of logistic regression (LR) with unigrams, bi-grams, and TF-IDF was suggested in [26] and subsequently applied to the BRAD dataset. Likewise, the LR technique was recommended in [27] using similar features and was later implemented in the HARD dataset. Our ST-SA model also underwent a comparative evaluation using the LABR datasets. This evaluation included reference methods, such as SVM, which utilized a support vector machine classifier with n-gram features as indicated in [23], MNB that applied a multinomial naive Bayes technique with bag-of-words features as per [22], and HC, a hierarchical classifiers model based on the divide-and-conquer method mentioned in [24]. Additionally, HC(KNN) represents a refined version of the hierarchical classifiers, maintaining its foundation in the divide-and-conquer approach, as described in [25]. Recently, significant advancements in natural language processing (NLP) have been made possible through the bidirectional encoder representations from transformers, or BERT [57]. Specifically, AraBERT [58], a BERT model pre-trained in Arabic, was trained on three diverse corpora, OSIAN [59], Arabic Wikipedia, and the MSA corpus, collectively amounting to approximately 1.5 billion words. We performed a comparative study between the proposed ST-SA architecture for Arabic dialects and other models, like AraBERT [58] and T-TC-INT [44], examining their effectiveness and utility in various contexts.

4.5. Results

Numerous empirical experiments were conducted employing the proposed ST-SA system for Arabic dialects. The suggested ST-SA system underwent training with varying configurations of attention heads (AHs) in the MHA sub-layer and diverse encoder quantities to ascertain the most efficient structure. Additionally, the system was trained with varying dimensions of word embeddings for each token. This research assessed the influence of training the proposed system using two multi-tasking methodologies, namely, in tandem and alternatively, for the performance assessment. The efficacy of the suggested system's sentiment analysis was assessed using an automated accuracy metric. This section details the evaluation of the proposed ST-SA system across five-polarity classification tasks for ADs. The results of the practical experiments on HARD, BRAD, and LABR are delineated in Tables 5–7 respectively. As elucidated in Figure 4, Tables 5 and 8, the proposed ST-SA system achieves an accuracy of 84.02%, F-score value of 83.50%, and precision value of 83.97% on the HARD imbalanced dataset, where the number of AH is 2, number of tokens is 90, number of experts is 10, batch size is 60, filter size is 32, dropout value is 0.25, and the embedding dimension for each token is 23. This commendable accuracy was achieved due to the favorable impact of employing the MTL framework, MoE mechanism, and MHA approach, particularly in right-to-left texts, like ADs. MoE employs a collection of expert networks to grasp distinct facets of the input data, subsequently amalgamating their outputs via a gating network. This enables the model to dynamically select from various parameter sets (i.e., expert modules) based on the input so that the proposed model can detect the sentiments accurately. When compared to the performance of the leading system on the HARD dataset, the results achieved by the ST-SA model exceeded

those attained by LR [60] by a margin of 7.92% in terms of the accuracy. Additionally, the proposed model outperformed AraBERT [58] with an accuracy difference of 3.17% and surpassed the T-TC-INT model [45] by 2.19% in terms of the accuracy. As a result, the concurrent execution of learning-related tasks expanded the available data pool and reduced the risk of overfitting [61]. The presented system exhibited proficiency in capturing both syntactic and semantic attributes, allowing it to accurately identify the sentiments conveyed in AD sentences.

Table 5. Results for the proposed ST-SA model on the HARD dataset for the five-polarity classification task, where E-D-T is the embedding dimension for each token, NT is the number of tokens, AH is the number of attention heads, FS is the filter size, NE is the number of experts, BS is the batch size, and DO is the dropout value.

E-D-T	NT	AH	FS	NE	BS	DO	Accuracy (5-Polarity)	F-Score	Precision
50	50	4	50	10	50	0.30	81.39%	80.79%	80.98%
32	100	2	32	10	50	0.25	83.81%	82.47%	83.08%
23	90	2	32	10	60	0.25	84.02%	83.50%	83.97%
30	150	4	30	5	50	0.25	82.89%	81.67%	82.18%
30	25	4	30	5	50	0.30	82.72%	80.16%	81.56%

Table 6. Results for the ST-SA model on the BRAD dataset for the five-polarity classification task.

E-D-T	NT	AH	FS	NE	BS	DO	Accuracy (5-Polarity)	F-Score	Precision
30	20	2	30	6	40	0.22	66.72%	65.53%	65.69%
40	15	3	30	10	55	0.25	67.37%	66.27%	66.48%
35	17	3	35	13	52	0.30	64.95%	63.76%	63.37%
50	24	3	30	15	53	0.24	68.81%	67.89%	68.13%
55	30	3	40	18	56	0.26	67.15%	66.80%	66.91%

Table 7. Results for the ST-SA model on the LABR dataset for the five-polarity classification task.

E-D-T	NT	AH	FS	NE	BS	DO	Accuracy (5-Polarity)	F-Score	Precision
40	20	3	35	10	50	0.30	80.09%	79.25%	79.87%
60	100	3	35	12	70	0.27	83.91%	82.71%	83.03%
35	40	2	40	10	60	0.20	81.74%	80.00%	80.24%
20	40	4	39	15	40	0.30	82.65%	80.17%	81.77%
30	40	4	40	12	50	0.30	81.49%	80.44%	80.82%

Table 8. The performance of the proposed ST-SA model compared with benchmark approaches on the HARD imbalanced dataset.

Model	Polarity	Accuracy	F-Score
LR [60]	5	76.1%	75.90%
AraBERT [58]	5	80.85%	77.88%
T-TC-INT [44]	5	81.83%	80.91%
Proposed ST-SA Model	5	84.02%	83.50%

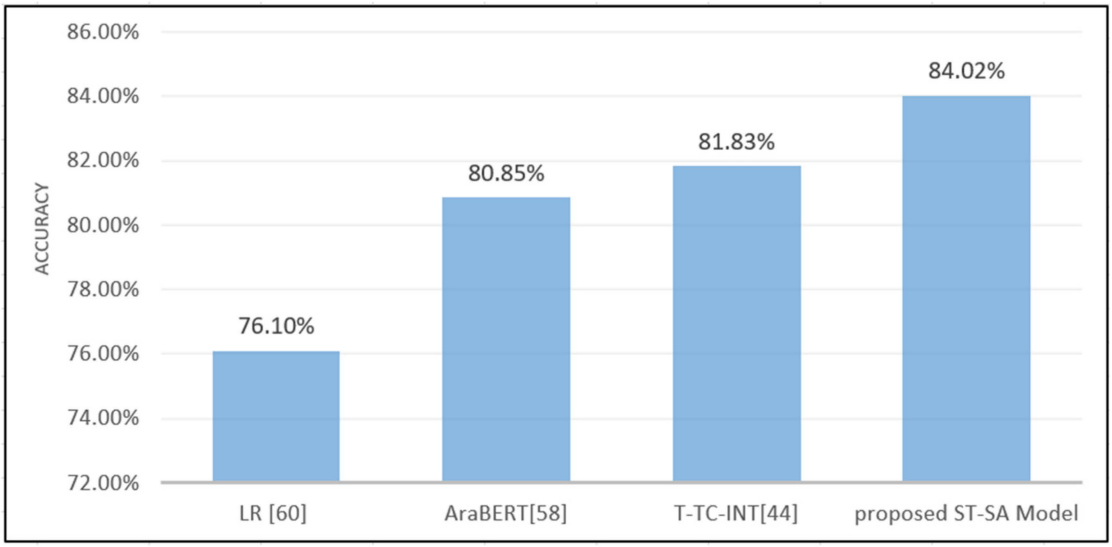

Figure 4. Evaluation accuracy of the proposed ST-SA model in comparison with state-of-the-art approaches on the HARD test dataset.

Furthermore, the recommended ST-SA system showcased a superior performance on the imbalanced BRAD dataset. As illustrated in Table 6, the proposed model achieves an accuracy rate of 68.81%, an F-score of 67.89%, and a precision score of 68.13%. These results were obtained when the number of AH was 3, the number of tokens was 24, the number of experts was 15, the batch size was 53, the filter size was 30, the dropout value was 0.24, and the embedding dimension for each token was 50. As detailed in Table 9 and Figure 5, the suggested ST-SA system outperforms the logistic regression (LR) approach advocated by [26] by a substantial margin, exhibiting an accuracy difference of 21.71%. Additionally, it surpasses the AraBERT model [58] by a margin of 7.96% and the T-TC-INT [44] system by 7.08%. In addition, the incorporation of the switch-transformer-based shared encoder, with one for each classification task, enabled the suggested model to capture a comprehensive representation that encompassed the preceding, subsequent, and localized contexts of any position within a sentence.

Table 9. The performance of the proposed ST-SA model compared with benchmark approaches on the BRAD imbalanced dataset.

Model	Polarity	Accuracy	F-Score
LR [26]	5	47.7%	48.90%
AraBERT [58]	5	60.85%	58.79%
T-TC-INT [44]	5	61.73%	61.40%
Proposed ST-SA Model	5	68.81%	67.89%

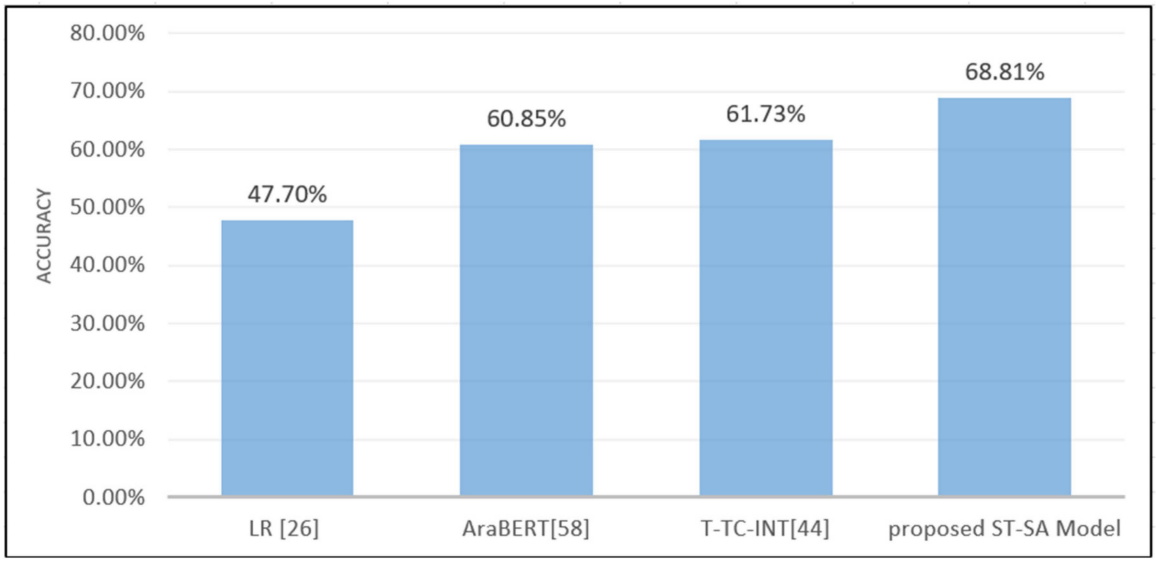

Figure 5. Evaluation accuracy of the proposed ST-SA model in comparison with state-of-the-art approaches on the BRAD test dataset.

Furthermore, the recommended switch-transformer sentiment analysis model with multi-task learning (ST-SA), as detailed in Table 7, demonstrated an outstanding performance on the challenging and imbalanced LABR dataset. In this study, this innovative model achieved a remarkable accuracy of 83.91%, an F-score of 82.71%, and a precision score of 83.03%, surpassing alternative approaches. It is worth noting that, with specific settings, including three attention heads (AHs), a filter size of 35, 100 tokens, 12 experts, a batch size of 70, a dropout value of 0.27, and an embedding dimension of 60 for each token, the suggested system truly showcases its effectiveness. This achievement highlights the robustness of the ST-SA model for addressing the challenges of sentiment analyses within the context of an imbalanced dataset. As demonstrated in Table 10 and Figure 6, the proposed switch-transformer sentiment analysis model with multi-task learning (ST-SA) exhibits superiority over various alternative methods. Notably, the ST-SA model outperformed multiple models by substantial margins. For example, it achieved a significant accuracy difference of 33.61% compared to the SVM [23] model, an impressive 38.91% accuracy difference surpassing the MNP [22] model, a substantial 26.11% accuracy difference over the HC(KNN) [24] model, as well as a noteworthy 24.95% accuracy difference when compared to AraBERT [58]. The proposed model even surpassed HC(KNN) [25] by an accuracy difference of 11.27%. Additionally, the proposed model outperformed the T-TC-INT model [44] with an accuracy difference of 5.78%.

Table 10. The performance of the proposed ST-SA model compared with benchmark approaches on the LABR imbalanced dataset.

Model	Polarity	Accuracy	F-Score
SVM [23]	5	50.3%	49.1%
MNP [22]	5	45.0%	42.8%
HC(KNN) [24]	5	57.8%	63.0%
AraBERT [58]	5	58.96%	55.88%
HC(KNN) [25]	5	72.64%	74.82%
T-TC-INT [44]	5	78.13%	77.80%
Proposed ST-SA Model	5	83.91%	82.71%

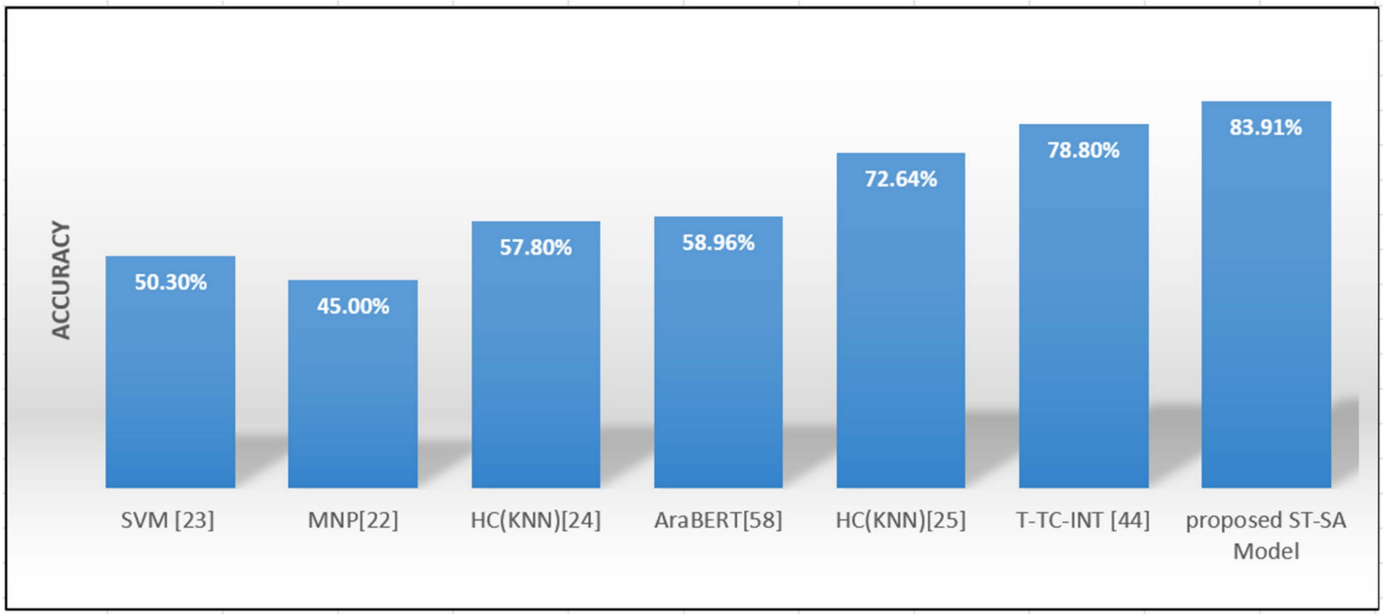

Figure 6. Evaluation accuracy of the proposed ST-SA model in comparison with state-of-the-art approaches on the LABR test dataset.

In the context of deep learning, joint training refers to simultaneously training a single neural network on multiple related tasks. Instead of creating separate models for each task, this method allows the model to learn and leverage common features across all tasks, improving its adaptability and efficiency. Often, this approach results in better performances for each task as the model can benefit from the interconnectedness of the tasks. Imbalanced data, on the other hand, refer to a dataset where the distribution of classes (or categories) is not uniform. Frequently, this means that one or several classes have significantly fewer instances than others, which can pose challenges in model training and performance evaluation.

This situation can present difficulties for deep learning models as they can exhibit a bias towards the majority class, resulting in subpar performances on minority classes. The evaluation results suggest that the presented ST-SA system, when subjected to both joint and alternate learning, exhibits exceptional efficiency. Alternate training demonstrated superior results compared to joint learning, achieving accuracies of 84.02% and 76.62% on the imbalanced HARD dataset, and 67.37% and 64.23% on BRAD, as detailed in Table 11. In comparison to the standard methods, alternate training in a five-point classification system appears to capture more nuanced feature representations in text sequences than singular task learning. These findings indicate that alternate learning is more effective for complex sentiment analysis (SA) tasks, capable of developing complex and more detailed, latent representations for Arabic dialect sentiment analysis (AD SA) tasks. The notable difference in performance between the two approaches is attributed to how alternate training benefits from the varying data volumes present in each task's dataset.

Table 11. Performance of joint and alternate training techniques for five-polarity classification.

ST-SA Training Method	HARD (Imbalance) Accuracy	BRAD (Imbalance) Accuracy
Alternately	84.02%	67.37%
Jointly	76.62%	64.23%

Shared layers tend to hold more information when a task involves a larger dataset. In contrast, joint learning can lean towards bias if one of the tasks is associated with a significantly larger dataset than the other. Consequently, alternative training methods are deemed more suitable for tasks related to the sentiment analysis of Arabic dialects.

This is particularly evident in cases involving two distinct datasets for separate tasks, such as machine translation tasks transitioning from Arabic dialects (ADs) to Modern Standard Arabic (MSA) and subsequently to English [2]. The performance of each task can be optimized by constructing the network in an alternating manner, thus eliminating the need for more training data [1]. Furthermore, leveraging synergies between related tasks can enhance the capabilities of five-point classification systems. The notable improvements in our model's performance can be attributed to various factors. Achieving superior results compared to established models, like AraBERT, known for its proficiency in Arabic language tasks, is an achievement in itself. By surpassing AraBERT on identical datasets, our model proves its superior accuracy in processing Arabic dialects. Even marginal increases in accuracy are significant as they contribute to the overall enhancement of models designed for Arabic dialect processing. These improvements can lead to practical benefits in several areas, including more precise sentiment analysis, improved information retrieval, and other natural language processing applications tailored to Arabic dialects.

Significantly, the ST-SA system did not demonstrate marked enhancements on the BRAD dataset relative to the current models. This could be due to the unique aspects, expressions, and linguistic nuances of the BRAD Arabic dataset, which the ST-SA sentiment analysis model might not fully comprehend. Without a proper domain adaptation, there is a potential mismatch between the model's acquired features and the specific traits of the BRAD dataset, leading to a less than optimal performance. Improving a model's performance on the BRAD sentiment analysis dataset using advanced deep learning approaches through domain adaptation involves several key techniques. For example, transformers, particularly BERT (bidirectional encoder representations from transformers), have revolutionized NLP by effectively capturing context in text. Fine-tuning a pre-trained BERT model on the BRAD dataset can significantly enhance the sentiment analysis performance.

For the feasibility assessment, pre-trained models are readily available, but fine-tuning requires substantial computational resources and expertise in NLP. It is feasible if these resources are accessible. Another key technique is adversarial training, which involves training the model to be robust against adversarial examples designed to deceive it. In the sentiment analysis of five-polarity ADs, this can make the model more resilient to nuances and variations in sentiment expressions. Implementing adversarial training can be complex and computationally intensive but is feasible with adequate resources and expertise in deep learning. Domain-adaptive fine-tuning is one the approaches and techniques that can present an outstanding performance regarding sentiment analyses for the BRAD dataset. This technique involves gradually fine-tuning a pre-trained model on a mix of source and target domain data, increasingly focusing on the target domain. This helps the model adapt to the specific language and sentiment expressions of the BRAD dataset. In addition, domain-adaptive fine-tuning is practical if there is enough data from both the source and target domains. It is less resource-intensive compared to training a model from scratch. Furthermore, meta-learning trains a model for a variety of tasks to learn how to adapt quickly to new tasks or domains. This approach is useful in five-polarity sentiment analyses for ADs for handling diverse expressions and contexts. Also, meta-learning requires diverse training datasets and significant computational power. It is feasible in well-resourced environments. If the BRAD dataset includes multilingual data, cross-lingual models, like multilingual BERT, can be effective. These models are trained in multiple languages and can handle sentiment analyses across different linguistic contexts. Similar to BERT, these models are available pre-trained. Fine-tuning on the specific languages of the BRAD dataset is necessary and feasible with appropriate computational resources.

4.6. Impact of Number of Experts (NE)

As demonstrated in Tables 5–7, the effectiveness of the recommended ST-SA framework across diverse input representations derived from the self-attention layer underscores the significance of the proposed model for the classification task encompassing five distinct polarities. Here, "NE" denotes the number of experts in the encoding layer within the

suggested switch-transformer SA model, employing the MoE mechanism. The devised system underwent training utilizing varying expert numbers: 5, 6, 10, 12, 13, 15, and 18. As evident in Tables 5–7, a discernible shift in the accuracy scores is observed for the HARD, BRAD, and LABR categories.

4.7. Impact of Length of Input Sentence

Acquiring extended syntactic dependencies and contextual comprehension across elements in input expressions enhances the efficacy of classifying lengthy sentences. Sentences of equivalent lengths (in terms of source tokens) were clustered together as demonstrated in the work of Luong et al. [62]. Due to the substantial scale of the HARD corpus, a task involving a five-fold classification of polarities on the HARD dataset was selected to assess the performance of the self-attention (SA) mechanism for long sentences. The assessment in this section was predicated on the subsequent ranges: <10, 10–20, 20–30, 30–40, 40–50, and >50. An automated accuracy measure was computed for the output generated by the switch-transformer sentiment analysis system. As depicted in Table 12, the effectiveness of the recommended switch-transformer sentiment analysis (ST-SA) model increased as the length of the input sentences extended. This improvement was especially noticeable when dealing with sentences consisting of 40- to 50-word tokens and those exceeding 50-word tokens, resulting in accuracy scores of 81.32% and 84.02%, respectively. Through the employment of multi-task learning, a multi-headed attention mechanism, mixture of experts (MoE) mechanism, and the incorporation of word units as an input characteristic for the MHA sub-layer, the proposed system attained contextually pertinent knowledge and dependencies of the tokens, regardless of their position in the AD input phrases. Furthermore, the utilization of MoE in the switch transformer enabled it to excel at discerning complex patterns within the input domain, leveraging the specialized knowledge of a multitude of experts. However, the efficiency of the suggested model was notably lower for shorter sentences, specifically those comprising 10- to 20-word tokens, 30- to 40-word tokens, and 20- to 30-word tokens. Furthermore, the system's effectiveness notably declined for sentences with fewer than 10-word tokens, yielding a meager accuracy of 77.25%. The impressive performance of the recommended ST-SA system across various sentence lengths underscores the efficacy of leveraging the MHA methodology and MTL framework, along with employing mixture of experts (MoE) mechanism, in enhancing the encoder's MHA sublayer proficiency in discerning word relationships within the AD input sentences.

Table 12. Accuracy score on HARD dataset with different sentence lengths.

Sentence Length	Accuracy
<10	77.25%
(10–20)	77.35%
(20–30)	77.95%
(30–40)	78.63%
(40–50)	81.32%
>50	84.02%

4.8. Motivation and Novelty

In the context of sentiment analyses for Arabic Dialects (ADs), our research introduced an innovative switch-transformer sentiment analysis (ST-SA) model tailored to the five-point categorization of ADs. The motivation behind this work stemmed from the need for effective sentiment analyses in a domain characterized by linguistic intricacies and limited training data. The novel aspects and motivations of this study can be summarized as follows:

1. Enhancing representation with multi-task learning (MTL): our approach incorporated a multi-task learning (MTL) framework, coupled with the self-attention mechanism, to enrich the representation of textual sequences. This novel combination aimed to

improve the system's ability to capture both global and local semantic knowledge within the context.

2. Effective handling of imbalanced data: we addressed the challenge of imbalanced data in sentiment analyses, particularly in the context of ADs. This was crucial for performing accurate sentiment classifications in a domain where certain sentiments could be less frequent.
3. Multi-head attention (MHA) strategy: the utilization of the multi-head attention (MHA) strategy within our model allowed for the identification of key terms and words within text sequences. This strategy significantly contributed to the model's proficiency in understanding the nuances of ADs.
4. Fine-grained sentiment analysis: we explored a fine-grained approach, including ternary classifications within the MTL framework, to further refine sentiment discrimination. This approach enhanced the differentiation between high-negative and negative categories within the five-point classification schema.
5. Superior performance: the empirical results demonstrate the superiority of our ST-SA model over existing state-of-the-art methodologies across multiple datasets, including HARD, BRAD, and LABR. This highlights the practical effectiveness of our approach in real-world applications.
6. Addressing syntactic complexities: our model not only handled limited training data effectively, but also adeptly addressed the syntactic complexities inherent in the free-format nature of AD phrases. This unique capability sets our ST-SA system apart.
7. Incorporating advanced techniques: the ST-SA system incorporated cutting-edge techniques, including the multi-head attention (MHA) strategy, mixture of experts (MoE) mechanism, and word units as input features. These innovations collectively resulted in a highly proficient sentence classification system tailored specifically for ADs.

4.9. Principal Findings

- The study proposed an innovative approach, the switch-transformer multi-task learning (ST-MTL) model, for classifying Arabic dialects (ADs) into five distinct categories. This method combined multi-task learning (MTL) with a cutting-edge switch-transformer model.
- The incorporation of a switch transformer and particularly the mixture of experts (MoE) mechanism served to augment the portrayal of the comprehensive text sequence on a global scale.
- The model's ability to harness insights from multiple experts, each specializing in distinct facets of the data, enabled it to amalgamate their discoveries, resulting in an enhancement of the overall performance.
- The proposed model offered a promising solution to address the constraints imposed by limited datasets in the context of text classification for advertisements (ADs).
- Elevating quality through MTL and switch transformer: combining the multi-task learning (MTL) framework with the inclusion of word units as input features for the MHA sub-layer in the switch-transformer encoder offered significant benefits.
- Superior performance of alternate learning over joint learning: the results indicate that opting for alternate learning, rather than joint learning, leads to enhanced effectiveness.
- Impact of input sentence length: the effectiveness of the recommended ST-SA model increased as the length of the input sentences extended, especially for sentences containing 40- to 50-word tokens achieving an accuracy score of 81.32%.
- The model excelled with sentences exceeding 50-word tokens, achieving a remarkable accuracy score of 84.02%.
- Cutting-edge advancement: the empirical findings from the practical experimentation of the suggested model clearly demonstrate its supremacy over current methodologies.
- The findings are supported by outstanding total accuracy rates: 84.02% for the HARD dataset, 68.81% for the BRAD dataset, and 83.91% for the LABR dataset.

- Notably, these results signify a substantial improvement when compared to well-known models, like T-TC-INT, AraBERT, and LR.

5. Conclusions

We introduced an ST-SA model designed for the five-point categorization of Arabic dialects (ADs). The proposed framework utilized the self-attention approach and incorporated a multi-task learning (MTL) framework to enhance the overall representation of the text sequence. Moreover, the MHA methodology was adept at singling out the most pertinent terms and words within the text sequences. Through training on sentiment analysis (SA) tasks encompassing ternary and five-polarity assignments for ADs, the system's effectiveness was notably enhanced. The utilization of MHA in conjunction with MTL markedly elevated the quality of the proposed SA system. The outcomes of this study underscore the pivotal attributes of the ST-SA system, which employs the MoE mechanism and MHA approach to augment the accuracies of both five-point and three-point classification tasks. The integration of the MTL framework, MoE mechanism, and word units as input characteristics in the MHA sub-layer underscores the critical role of these strategies in low-resource language SA tasks, such as ADs. Similarly, experimenting with various configurations, including the deployment of multiple heads in the MHA sub-layer and training with multiple numbers of experts empowers the proposed ST-SA to master intricate patterns within the input domain by capitalizing on the specialized expertise of numerous experts. Also, it led to a notable boost in the classification performance of the proposed system. Conducting a series of experiments on two datasets for five-point Arabic SAs, our findings reveal that alternate learning paradigms demonstrate superior efficiency compared to joint learning, with the dataset size of each task exerting an influence. The outcomes clearly reveal that the suggested system surpasses other advanced methods when tested on the HARD, BRAD, and LABR datasets. Additionally, it was observed that employing alternate training for tasks within the model based on the multi-task learning (MTL) framework could considerably improve the performance of the five-point classification. Specifically, adopting a detailed ternary classification strategy, especially in identifying text as negative, aided in more accurately differentiating between the high-negative and negative categories within the five-point classification structure.

Practical experiments on five-point and three-point categorization tasks demonstrated that the recommended system significantly improved the accuracy compared to other sentiment analysis systems for Arabic dialects. The proposed switch-transformer sentiment analysis (ST-SA) system that utilized MTL generated a resilient latent feature representation for textual sequences in Arabic dialects. With overall accuracy rates of 84.02%, 68.81%, and 83.91% for the HARD, BRAD, and LABR datasets, correspondingly, the empirical findings underscore the superior performance of the ST-SA model over existing state-of-the-art methodologies. Examples include Ar-aBERT [58], support vector machine (SVM) [23], multi-neural perceptron (MNP) [22], hierarchical clustering with k-nearest neighbor (HC(KNN)) [24], and logistic regression (LR) [26], and T-TC-INT [44]. Further analyses of the experiments and outcomes unveiled that the system's efficacy was contingent on the utilization of the multi-head attention (MHA) strategy and the dimensionality of word embeddings for each token. The practical investigation elucidated the advantages of employing the MHA technique, as it enabled the extraction of both global and local semantic knowledge within the contextual framework through the MHA sub-layer in each encoding layer.

In addition, the proposed ST-SA (sentiment analysis for ADs) system not only addressed the challenge of limited training data, but also adeptly tackled the syntactic complexities inherent in the free-format nature of AD (advertisement) phrases. This unique approach sets the ST-SA system apart as it incorporates cutting-edge techniques, including the multi-head attention (MHA) strategy, mixture of experts (MoE) mechanism, and word units used as input features for the MHA sub-layer. These innovations collectively result in a highly proficient sentence classification system tailored specifically for ADs, allow-

ing for an accurate sentiment analysis in this domain. Looking ahead, our future plans include further enhancing the ST-SA system's capabilities. We are actively working on the development of a multi-task learning sentiment analysis architecture that leverages sub-word units as input features for the MHA sub-layer, as recommended by the recent research [52]. Furthermore, we are exploring the adoption of a novel reverse positional encoding mechanism [63] to effectively address the syntactic and semantic intricacies frequently encountered in right-to-left textual content, such as ADs, and perform interpretability analysis, such as attention visualization. These advancements aim to reinforce the system's ability to handle diverse linguistic nuances and improve the accuracy of sentiment analysis, making it a valuable tool for analyzing sentiment in ADs across various contexts and languages.

Author Contributions: L.H.B. and S.K. conceived and designed the methodology and experiments; L.H.B. performed the experiments; L.H.B. analyzed the results; L.H.B. and S.K. analyzed the data; L.H.B. wrote the paper. S.K. reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT under Grant NRF-2022R1A2C1005316.

Data Availability Statement: The dataset generated during the current study is available from the [ST_SA_AD] repository (https://github.come/laith85, accessed on 1 January 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Baniata, L.H.; Park, S.; Park, S.-B. A Neural Machine Translation Model for Arabic Dialects That Utilizes Multitask Learning (MTL). *Comput. Intell. Neurosci.* **2018**, *2018*, 7534712. [CrossRef] [PubMed]
2. Baniata, L.H.; Park, S.; Park, S.-B. A multitask-based neural machine translation model with part-of-speech tags integration for Arabic dialects. *Appl. Sci.* **2018**, *8*, 2502. [CrossRef]
3. Salloum, S.A.; AlHamad, A.Q.; Al-Emran, M.; Shaalan, K. A survey of Arabic text classification. *Intell. Nat. Lang. Process. Trends Appl.* **2018**, *8*, 4352–4355.
4. Harrat, S.; Meftouh, K.; Smaili, K. Machine translation for Arabic dialects (survey). *Inf. Process. Manag.* **2019**, *56*, 262–273. [CrossRef]
5. El-Masri, M.; Altrabsheh, N.; Mansour, H. Successes and challenges of Arabic sentiment analysis research: A literature review. *Soc. Netw. Anal. Min.* **2017**, *7*, 54. [CrossRef]
6. Elnagar, A.; Yagi, S.M.; Nassif, A.B.; Shahin, I.; Salloum, S.A. Systematic Literature Review of Dialectal Arabic: Identification and Detection. *IEEE Access* **2021**, *9*, 31010–31042. [CrossRef]
7. Abdul-Mageed, M. Modeling Arabic subjectivity and sentiment in lexical space. *Inf. Process. Manag.* **2019**, *56*, 308–319. [CrossRef]
8. Al-Smadi, M.; Al-Ayyoub, M.; Jararweh, Y.; Qawasmeh, O. Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' reviews using morphological, syntactic and semantic features. *Inf. Process. Manag.* **2019**, *56*, 308–319. [CrossRef]
9. Baly, R.; Badaro, G.; El-Khoury, G.; Moukalled, R.; Aoun, R.; Hajj, H.; El-Hajj, W.; Habash, N.; Shaban, K.; Diab, M.; et al. A Characterization Study of Arabic Twitter Data with a Benchmarking for State-of-the-Art Opinion Mining Models. In Proceedings of the Third Arabic Natural Language Processing Workshop, Valencia, Spain, 3 April 2017; pp. 110–118.
10. El-Beltagy, S.R.; El Kalamawy, M.; Soliman, A.B. NileTMRG at SemEval-2017 Task 4: Arabic Sentiment Analysis. In Proceedings of the 11th International Workshop on Semantic Evaluation (semEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 790–795.
11. Jabreel, M.; Moreno, A. SiTAKA at SemEval-2017 Task 4: Sentiment Analysis in Twitter Based on a Rich set of Features. In Proceedings of the 11th International Workshops on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 692–697.
12. Mulki, H.; Haddad, H.; Gridach, M.; Babaoğlu, I. Tw-StAR at SemEval-2017 Task 4: Sentiment Classification of Arabic Tweets. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 664–669.
13. Siddiqui, S.; Monem, A.A.; Shaalan, K. Evaluation and enrichment of Arabic sentiment analysis. *Intell. Nat. Lang. Process. Trends Appl.* **2017**, *740*, 17–34.
14. Al-Azani, S.; El-Alfy, E.S. Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment analysis in short Arabic text. *Procedia Comput. Sci.* **2017**, *109*, 359–366. [CrossRef]
15. Alali, M.; Sharef, N.M.; Hamdan, H.; Murad, M.A.A.; Husin, N.A. Multi-layers convolutional neural network for twitter sentiment ordinal scale classification. *Adv. Intell. Syst. Comput.* **2018**, *700*, 446–454.

16. Alali, M.; Sharef, N.M.; Murad, M.A.A.; Hamdan, H.; Husin, N.A. Narrow Convolutional Neural Network for Arabic Dialects Polarity Classification. *IEEE Access* **2019**, *7*, 96272–96283. [CrossRef]
17. Gridach, M.; Haddad, H.; Mulki, H. Empirical evaluation of word representations on Arabic sentiment analysis. *Commun. Comput. Inf. Sci.* **2018**, *782*, 147–158.
18. Al Omari, M.; Al-Hajj, M.; Sabra, A.; Hammami, N. Hybrid CNNs-LSTM Deep Analyzer for Arabic Opinion Mining. In Proceedings of the 2019 6th International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019; pp. 364–368.
19. Fedus, W.; Zoph, B.; Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **2021**, *23*, 5232–5270.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–9008.
21. Jin, N.; Wu, J.; Ma, X.; Yan, K.; Mo, Y. Multi-task learning model based on multi-scale cnn and lstm for sentiment classification. *IEEE Access* **2020**, *8*, 77060–77072. [CrossRef]
22. Aly, M.; Atiya, A. LABR: A large scale Arabic book reviews dataset. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, 4–9 August 2013; Volume 2, pp. 494–498.
23. Al Shboul, B.; Al-Ayyoub, M.; Jararweh, Y. Multi-way sentiment classification of Arabic reviews. In Proceedings of the 2015 6th International Conference on Information and Communication Systems (ICICS), Amman, Jordan, 7–9 April 2015; pp. 206–211.
24. Al-Ayyoub, M.; Nuseir, A.; Kanaan, G.; Al-Shalabi, R. Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 531–539. [CrossRef]
25. Nuseir, A.; Al-Ayyoub, M.; Al-Kabi, M.; Kanaan, G.; Al-Shalabi, R. Improved hierarchical classifiers for multi-way sentiment analysis. *Int. Arab J. Inf. Technol.* **2017**, *14*, 654–661.
26. Elnagar, A.; Einea, O. BRAD 1.0: Book reviews in Arabic dataset. In Proceedings of the 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), Agadir, Morocco, 29 November–2 December 2016.
27. Elnagar, A.; Khalifa, Y.S.; Einea, A. Hotel Arabic-reviews dataset construction for sentiment analysis applications. *Stud. Comput. Intell.* **2018**, *740*, 35–52.
28. Balikas, G.; Moura, S.; Amini, M.-R. Multitask Learning for Fine-Grained Twitter Sentiment Analysis. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, 7–11 August 2017; pp. 1005–1008.
29. Lu, G.; Zhao, X.; Yin, J.; Yang, W.; Li, B. Multi-task learning using variational auto-encoder for sentiment classification. *Pattern Recognit. Lett.* **2020**, *132*, 115–122. [CrossRef]
30. Sohangir, S.; Wang, D.; Pomeranets, A.; Khoshgoftaar, T.M. Big Data: Deep Learning for financial sentiment analysis. *J. Big Data* **2018**, *5*, 3. [CrossRef]
31. Jangid, H.; Singhal, S.; Shah, R.R.; Zimmermann, R. Aspect-Based Financial Sentiment Analysis using Deep Learning. In Proceedings of the Companion of the Web Conference 2018 on The Web Conference, Lyon, France, 23–27 April 2018; pp. 1961–1966.
32. Ain, Q.T.; Ali, M.; Riaz, A.; Noureen, A.; Kamran, M.; Hayat, B.; Rehman, A. Sentiment analysis using deep learning techniques: A review. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 424.
33. Gao, Y.; Rong, W.; Shen, Y.; Xiong, Z. Convolutional neural network based sentiment analysis using Adaboost combination. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 1333–1338.
34. Hassan, A.; Mahmood, A. Deep learning approach for sentiment analysis of short texts. In Proceedings of the Third International Conference on Control, Automation and Robotics (ICCAR), Nagoya, Japan, 24–26 April 2017; pp. 705–710.
35. Qian, J.; Niu, Z.; Shi, C. Sentiment Analysis Model on Weather Related Tweets with Deep Neural Network. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Macau, China, 26–28 February 2018; pp. 31–35.
36. Pham, D.-H.; Le, A.-C. Learning multiple layers of knowledge representation for aspect based sentiment analysis. *Data Knowl. Eng.* **2018**, *114*, 26–39. [CrossRef]
37. Preethi, G.; Krishna, P.V.; Obaidat, M.S.; Saritha, V.; Yenduri, S. Application of deep learning to sentiment analysis for recommender system on cloud. In Proceedings of the 2017 International Conference on Computer, Information and Telecommunication Systems (CITS), Dalian, China, 21–23 July 2017; pp. 93–97.
38. Alharbi, A.S.M.; de Doncker, E. Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. *Cogn. Syst. Res.* **2019**, *54*, 50–61. [CrossRef]
39. Abid, F.; Alam, M.; Yasir, M.; Li, C.J. Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter. *Future Gener. Comput. Syst.* **2019**, *95*, 292–308. [CrossRef]
40. Wang, B.; Dong, G.; Zhao, Y.; Li, R. Learning from Fourier: Leveraging Frequency Transformation for Emotion Recognition. In *International Conference on Neural Information Processing*; Springer International Publishing: Cham, Switzerland, 2022.
41. Wang, B.; Dong, G.; Zhao, Y.; Li, R.; Yang, H.; Yin, W.; Liang, L. Spiking Emotions: Dynamic Vision Emotion Recognition Using Spiking Neural Networks. In Proceedings of the 2nd International Conference on Algorithms, High Performance Computing and Artificial Intelligence, Guangzhou, China, 21–23 October 2022.

42. Wang, B.; Dong, G.; Zhao, Y.; Li, R.; Cao, Q.; Hu, K.; Jiang, D. Hierarchically stacked graph convolution for emotion recognition in conversation. *Knowl.-Based Syst.* **2023**, *263*, 110285. [CrossRef]
43. Wang, B.; Dong, G.; Zhao, Y.; Li, R.; Cao, Q.; Chao, Y. Non-uniform attention network for multi-modal sentiment analysis. In *International Conference on Multimedia Modeling*; Springer International Publishing: Cham, Switzerland, 2022; pp. 612–623.
44. Baniata, L.H.; Kang, S. Transformer Text Classification Model for Arabic Dialects That Utilizes Inductive Transfer. *Mathematics* **2023**, *11*, 4960. [CrossRef]
45. Alali, M.; Mohd Sharef, N.; Azmi Murad, M.A.; Hamdan, H.; Husin, N.A. Multitasking Learning Model Based on Hierarchical Attention Network for Arabic Sentiment Analysis Classification. *Electronics* **2022**, *11*, 1193. [CrossRef]
46. Singh, S.; Kaur, H.; Kanozia, R.; Kaur, G. Empirical Analysis of Supervised and Unsupervised Machine Learning Algorithms with Aspect-Based Sentiment Analysis. *Appl. Comput. Syst.* **2023**, *28*, 125–136. [CrossRef]
47. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111–132. [CrossRef]
48. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
49. Xue, F.; Shi, Z.; Wei, F.; Lou, Y.; Liu, Y.; You, Y. Go wider instead of deeper. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 8779–8787.
50. Lazaridou, A.; Kuncoro, A.; Gribovskaya, E.; Agrawal, D.; Liska, A.; Terzi, T.; Gimenez, M.; de Masson d'Autume, C.; Kocisky, T.; Ruder, S.; et al. Mind the gap: Assessing temporal generalization in neural language models. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29348–29363.
51. Fan, A.; Bhosale, S.; Schwenk, H.; Ma, Z.; El-Kishky, A.; Goyal, S.; Baines, M.; Celebi, O.; Wenzek, G.; Chaudhary, V.; et al. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.* **2021**, *22*, 4839–4886.
52. Baniata, L.H.; Ampomah, I.K.E.; Park, S. A Transformer-Based Neural Machine Translation Model for Arabic Dialects that Utilizes Subword Units. *Sensors* **2021**, *21*, 6509. [CrossRef] [PubMed]
53. Dean, J.; Monga'TensorFlow, R. Large-Scale Machine Learning on Heterogeneous Distributed Systems'. 2015. Available online: https://www.tensorflow.org/ (accessed on 1 June 2023).
54. Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd.: Birmingham, UK, 2017.
55. Varoquaux, G.; Buitinck, L.; Louppe, G.; Grisel, O.; Pedregosa, F.; Mueller, A. Scikit-learn: Machine Learning in Python. *GetMobile Mob. Comput. Commun.* **2015**, *19*, 29–33. [CrossRef]
56. Baziotis, C.; Pelekis, N.; Doulkeridis, C. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 747–754.
57. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
58. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the LREC 2020 Workshop Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 9–15.
59. Zeroual, I.; Goldhahn, D.; Eckart, T.; Lakhouaja, A. OSIAN: Open Source International Arabic News Corpus—Preparation and Integration into the CLARIN-infrastructure. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 28 July–2 August 2019; pp. 175–182.
60. Pang, B.; Lee, L. *Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval*; Now Publishers: Boston, MA, USA, 2008; pp. 1–135.
61. Liu, S.; Johns, E.; Davison, A.J. End-to-end multi-task learning with attention. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1871–1880.
62. Luong, M.-T.; Pham, H.; Manning, C.D. Effective approaches to attention-based neural machine translation. In Proceedings of the Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1412–1421.
63. Baniata, L.H.; Kang, S.; Ampomah, I.K.E. A Reverse Positional Encoding Multi-Head Attention-Based Neural Machine Translation Model for Arabic Dialects. *Mathematics* **2022**, *10*, 3666. [CrossRef]