Preprints.org

Article

Not peer-reviewed version

Switching Self-Attention Text Classification Model with Innovative Reverse Positional Encoding for Right-To-Left Languages: A Focus on Arabic Dialects

Laith H. Baniata * and Sangwoo Kang *

Posted Date: 6 February 2024

doi: 10.20944/preprints202402.0332.v1

Keywords: Switching Self-Attention; Reverse Positional Encoding (RPE) mothed; Text Classification (SA); Right-to-Left Text; five-polarity; ITL

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Switching Self-Attention Text Classification Model with Innovative Reverse Positional Encoding for Right-To-Left Languages: A Focus on Arabic Dialects

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: Transformer models have emerged as frontrunners in the field of natural language processing, primarily due to their adept use of self-attention mechanisms to grasp the semantic linkages between words in sequences. Despite their strengths, these models often face challenges in single-task learning scenarios, particularly when it comes to delivering top-notch performance and crafting strong latent feature representations. This challenge is more pronounced in the context of smaller datasets and is particularly acute for under-resourced languages such as Arabic. In light of these challenges, this study introduces a novel methodology for text classification of Arabic texts. This method harnesses the newly developed Reverse Positional Encoding (RPE) technique. It adopts inductive-transfer learning (ITL) framework combined with a switching self-attention shared encoder, thereby increasing the model's adaptability and improving its sentence representation accuracy. The integration of Mixture of Experts (MoE) and Reverse Positional Encoding (RPE) techniques empowers the model to process longer sequences more effectively. This enhancement is notably beneficial for Arabic text classification, adeptly supporting both the intricate five-point and the simpler ternary classification tasks. The empirical evidence points to its outstanding performance, achieving accuracy rates of 87.20% for the HARD dataset, 72.17% for the BRAD dataset, and 86.89% for the LABR dataset, as evidenced by the assessments conducted on these datasets.

Keywords: Switching Self-Attention; Reverse Positional Encoding (RPE) mothed; Text Classification (SA); Right-to-Left Text; five-polarity; ITL

MSC: 68T07

---

1. Introduction

Text classification involves the algorithmic identification and interpretation of the sentiment expressed in text-based content, including sentences, entire documents, or social media updates. This process enables companies to gather valuable feedback on public perception regarding their brands, products, or services by analyzing user-generated content from online interactions. Platforms like Twitter see a vast daily volume of content in Arabic and various Arabic dialects, a trend that's expected to persist with the continuous rise of user-generated content. Arabic expressions represent a significant portion, approximately five percent, of the global online discourse. Furthermore, the prominence of the Arabic language in the digital realm has surged notably in recent years. Arabic is a global language, spoken by over five hundred million people around the world and falls under the category of semantic languages. As the official language in over 21 countries, stretching from the Arabian Gulf to the Atlantic Ocean, Arabic is known for its linguistic richness and complexity. This complexity is particularly pronounced when compared to languages like English, owing primarily to the vast array of dialects it encompasses. Additionally, the distinct variations between Modern Standard Arabic (MSA) and the various Arabic dialects (ADs) introduce further intricacy to the language. Moreover, the usage of the Arabic language is characterized by widespread diglossia. This

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

indicates that local Arabic dialects are predominantly used in casual conversations, while Modern Standard Arabic (MSA) is reserved for formal or professional contexts. For example, Libyans may alternate between MSA and their indigenous dialects based on the situation. The Libyan dialect reflects the country's rich history, cultural essence, heritage, and collective experiences. Arabic dialects vary significantly across regions, including Levantine (spanning Palestine, Jordan, Syria, and Lebanon), Maghrebi (covering Morocco, Algeria, Libya, and Tunisia), Iraqi, Nile Basin dialects (in Egypt and Sudan), and those from the Arabian Gulf (across the UAE, Saudi Arabia, Qatar, Kuwait, Yemen, Bahrain, and Oman). The task of identifying sentiment-laden words within this vast linguistic diversity is challenging due to the intricate structures, orthography, and complexity of the language. Each Arabic-speaking nation has its own distinctive vernacular, adding to the linguistic complexity. For instance, Arabic content on social platforms often blends Modern Standard Arabic (MSA) with local dialects, leading to varied interpretations of the same term.

Additionally, a notable syntactic complexity in Arabic dialects (ADs) relates to the structure of sentences, especially the placement of verbs, subjects, and objects. A comprehension of this arrangement is pivotal for analyzing AD sentences. As indicated in literature surveys, languages fall into various structural categories such as subject–object–verb (as in Korean), subject–verb–object (like in English), and verb–object–subject, which is typical for Arabic [1]. Furthermore, certain languages, including ADs, exhibit variable word orders [2], adding depth to the understanding of subjects, objects, and other elements in AD expressions. Consequently, relying solely on a single-task learning approach and manual feature creation proves inadequate for text classification in Arabic dialects. Moreover, these variations in ADs pose considerable challenges for conventional deep learning models, as longer AD phrases introduce complex and nuanced context involving the object, verb, and subject. A common limitation of traditional deep learning techniques is their diminishing efficiency in handling extended input sequences, leading to a compromise in the text classification (TC) model's performance as the sequence lengthens. 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). The lack of uniform orthographic rules is a significant obstacle in Arabic dialects (ADs), marked by the dialect-specific use of prefixes and suffixes not present in Modern Standard Arabic (MSA). Moreover, the ambiguity of many Arabic words, whose meanings can vary based on diacritics in the same syntactic framework, adds to the complexity. Building effective deep learning-based text classification (TC) models for ADs is further complicated by the need for extensive training data, which is notably difficult to gather for these dialects, often described as unstructured and lacking in resources. This scarcity makes information retrieval particularly challenging [3]. As the volume of training data for ADs reduces, so does the precision of classification. Additionally, most tools designed for MSA do not accommodate the distinct nuances of ADs [4]. It's also crucial to note that depending solely on lexical aids, such as lexicons, may not be entirely effective for Arabic SA. The sheer diversity of words across different dialects means it's unlikely for a single lexicon to cover all variations [5]. Moreover, developing tools and resources specifically for ADs is an extensive and time-consuming process [6]. Recently, text classification in the Arabic language has become a focal point of research, primarily aimed at categorizing views and tweets to discern both binary and ternary emotional sentiments. Many of these methodologies [7–12] utilize lexicons and tweet-specific attributes as inputs for machine learning (ML) algorithms. Conversely, some strategies adopt a rule-based framework, implementing lexicalization principles. This process involves formulating a hierarchy of heuristic rules to accurately classify tweets as negative or positive sentiments [13]. On another note, the Arabic sentiment ontology introduces a spectrum of sentiment intensities to distinguish user attitudes and facilitate the tweet classification process. Deep learning methods for text classification, including RNNs [14], CNNs [15–18], and recursive auto-encoders, have garnered significant attention due to their inherent adaptability and strength in automatic feature identification. Particularly, the advanced Switching Self-Attention model [19] has demonstrated superior performance over traditional transformer models [20] and RNN-based models in various natural language processing (NLP) tasks, thus drawing the interest of scholars in the deep learning domain.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

The Switching Self-Attention architecture capitalizes on a self-attention mechanism to analyze the connections between word pairs in a sentence. This model significantly boosts the efficacy of a variety of Natural Language Processing (NLP) tasks. It's particularly noteworthy in text classification, where it consistently proves to be the optimal choice across diverse datasets. The model's encoder uses self-attention to zero in on words within individual sentences. Furthermore, the Switching Self-Attention adeptly captures each word's positional data, like their order in the sentence, through positional encoding (PE), marking a departure from traditional RNN and CNN methodologies. The design of the Switching Self-Attention is underpinned by Multi-Head Attention (MHA) layers, enabling it to adeptly manage sequences of different lengths.

This study introduces a Switching Self-Attention text classification model that employs the Reverse Positional Encoding (RPE) technique to simplify the task into smaller, more digestible sub-tasks. Within each expert layer, the router determines which expert is best suited to process the token, basing its decision on the specific attributes of the token's representation. While the router effectively matches the token to the most appropriate expert based on its current representation, it doesn't consider how other experts might have handled the token. The model excels in managing extensive sequences and complex input-output dynamics, enhancing performance in both five-point and ternary Arabic text classification tasks. Although previous initiatives have tackled the intricacies of Arabic Dialect text classification (ADs TC), the Inductive Transfer Learning (ITL) strategy has proven to be a particularly effective approach.

Inductive Transfer Learning

Inductive-Transfer Learning (ITL) in deep learning is an influential method designed to boost efficiency and accuracy in learning by training a model on several related tasks at once. This strategy capitalizes on the similarities and variances among tasks, enhancing the model's ability to generalize effectively for each one. In ITL, task-shared representations allow the model to make use of valuable information across related tasks, which helps in minimizing overfitting risks associated with any single task. This approach is especially advantageous when data availability is limited for certain tasks. ITL models typically employ shared layers to acquire common attributes while using task-specific layers for unique features of each task, leading to stronger, more versatile models. Consequently, ITL is applied in various domains, such as natural language processing, where a single model can concurrently learn multiple tasks like text classification, language translation, and named entity recognition, benefiting from the interconnected nature of these tasks to improve overall performance. Inductive-Transfer Learning (ITL) boosts comprehension, optimizes encoder performance, and heightens the precision of sentiment classification relative to conventional single-task classifiers. This is achieved through the concurrent management of associated tasks and the adoption of a cohesive representation of text sequences [21]. A key advantage of Inductive-Transfer Learning (ITL) is its ability to effectively leverage varied resources for analogous tasks. However, it's important to note that most current methods for text classification (TC) of Arabic Dialects (ADs) primarily concentrate on binary and ternary classifications. In our research, we shift focus to explore the less examined area of five-polarity ADs TC. Uniquely, the application of a Switching Self-Attention framework along with ITL and Reverse Positional Encoding (RPE) technique for ADs Text classification is a novel approach not previously investigated. Earlier strategies in this field have generally depended on traditional transformers and Bi-LSTM techniques. Our contributions in this study can be summarized as follows:

- This study presents an innovative text classification model, termed Switching Self-Attention (SSA-TC-RPE), which leverages reverse positional encoding tailored for processing right-to-left texts like Arabic Dialects (ADs). The proposed SSA-TC-RPE model, utilizing this reverse positional encoding technique, demonstrates enhanced performance and effectiveness in analyzing sentiments in ADs, outperforming models that rely on extensive vocabularies.
- The text classification model, featuring Switching Self-Attention, incorporates the Mixture of Experts (MoE) strategy. This technique dissects the overarching task into smaller, more

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

controllable units, thereby enhancing the model's proficiency in effectively handling and interpreting sequences of greater length.

- Furthermore, the multi-head attention (MHA) mechanism is integrated with a Switching Self-Attention encoder to effectively manage ternary and five-category classifications in Arabic dialect texts. This method, part of a inductive transfer learning framework, aims to enhance text representation and broaden feature extraction.
- Moreover, the study unveils a novel approach that highlights the advantages of applying reverse positional encoding to right-to-left scripts, specifically Arabic dialects. This method's effectiveness is evaluated against traditional models that utilize standard absolute positional encoding for such text orientations.
- This study explores the impact of training the Switching Self-Attention model under varied conditions, including different embedding dimensions per token, a variety of token values, an assortment of attention head counts, multiple filter sizes, a spectrum of expert counts, a range of batch sizes, and various dropout values.
- The proposed SSA-TC-RPE model introduced in this study utilizes a Multi-Head Attention (MHA) mechanism to assess the relationship intensity between pairs of words in a sentence. This significantly enhances the effectiveness and significance of different tasks in the field of natural language processing.

The structure of this paper is laid out as follows: Section 2 presents a review of related literature, Section 3 details the methodology of the proposed model, Section 4 discusses the findings from our experiments, and Section 5 encapsulates the key takeaways and conclusions of the research.

2. Literature Review

Studies on Arabic text classification, especially those focusing on five-level polarity tasks, have been less prominent compared to binary or ternary classifications. Moreover, the prevailing methods tackling these classifications have predominantly utilized conventional machine learning techniques. For instance, research involving Arabic book reviews [22] investigated the use of corpora and lexicons, employing bag of words (BoW) features with various algorithms like passive aggressive (PA), support vector machine (SVM), logistic regression (LR), naive Bayes (NB), perceptron, and stochastic gradient descent (SGD). Another study [23] delved into the effects of stemming and balancing BoW features using similar machine learning algorithms on the same dataset, observing a decrease in performance with stemming. Research in [24] suggested a divide-and-conquer strategy for ordinal-scale classification tasks, utilizing a hierarchical classifier (HC) that segmented the five labels into more manageable sub-problems, showing that HC outperformed a singular classifier model. Building upon this concept, different hierarchical classifier architectures were proposed [25], comparing these with classifiers like SVM, KNN, NB, and DT. The findings indicated a boost in performance using hierarchical classifiers, although it's important to note that some of these models also experienced performance drops.

Another study [26] explored various machine learning classifiers, such as LR, SVM, and PA, applying n-gram features to analyze book reviews in the Arabic dataset (BRAD). The findings highlighted the superior performances of SVM and LR. In a related vein, research in [27] evaluated several sentiment classifiers, including AdaBoost, SVM, PA, random forest, and LR, on the hotel Arabic reviews dataset (HARD). This study found that SVM and LR, particularly with n-gram features, outperformed the others. These studies highlight a notable absence of deep learning techniques in classifying five levels of polarity in Arabic text classification (TC). Most existing methods for these five-level polarity tasks primarily use traditional machine learning algorithms, dependent on feature engineering, which is often time-intensive and complex. Additionally, these methodologies are based on single-task learning (STL) and do not adequately address the interconnectedness of different tasks (cross-task transfer) or simultaneously model various polarities, including both five and three levels. Several studies have employed Inductive Transfer (INT) to address the complexities of five-point text classification (TC) classification tasks. One such example is [28], where a framework based on inductive transfer learning was developed, employing a

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

recurrent neural network (RNN) to simultaneously handle both five-point and ternary classifications. This model integrated bidirectional long short-term memory (Bi-LSTM) and multilayer perceptron (MLP) layers. Furthermore, it enhanced feature extraction by including tweet-specific attributes such as counts of punctuation, elongated words, emoticons, and sentiment lexicons.

The outcomes of their study suggest that simultaneously training various text classification tasks greatly enhances the performance of the five-polarity task. Likewise, research in [29] capitalized on the combined potential of five-polarity and binary sentiment classification tasks through joint training. Their developed model featured an LSTM encoder and a variational auto-encoder decoder, serving as shared elements for both tasks. The findings demonstrated that the inductive-transfer learning (INT) model notably improved results for the five-polarity task. The advent of adversarial multi-task learning (AMTL) was first seen in [21], where the model combined two LSTM layers as task-specific elements with a single LSTM layer shared among tasks. In addition, the model incorporated a convolutional neural network (CNN) alongside the LSTM, and the outputs from both were merged with the output from the shared layer to create a comprehensive latent sentence representation. The researchers noted that their inductive transfer learning (ITL) model not only improved the outcomes of five-polarity classification tasks but also upgraded the quality of the encoder. While inductive transfer learning (INT) models have been applied in English language contexts, there's a significant gap in the application of multi-task and deep learning approaches for five-polarity classification in Arabic text.

Current research in this area mainly depends on single-task learning using conventional machine learning techniques. Therefore, there is considerable scope to advance the efficiency of Arabic text classification strategies, especially in handling the five levels of sentiment, where progress remains limited.

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.

Baniata et al. [44] presented a cutting-edge technique that employed an inductive transfer learning multi-head attention model for categorizing ADs into five levels. This advanced design combines a self-attention mechanism with an Inductive Transfer Learning (ITL) framework to enhance text sequence representation. Additionally, the self-attention mechanism is instrumental in isolating the most relevant words and phrases within these sequences. The model's performance saw a considerable boost when it was trained on text classification (TC) tasks, including ternary and five-polarity tasks tailored for ADs. The synergy between self-attention and ITL significantly improved the effectiveness of the proposed SA system. The results of this research highlight the key features of the ITL self-attention SA system, especially its use of self-attention to improve the accuracy of both five-point and three-point classification tasks. The integration of the ITL framework and word units as input for the self-attention sub-layer demonstrates their essential role in text classification tasks for low-resource languages like ADs. Furthermore, optimizing the model with various configurations, such as the use of multiple heads in the self-attention sub-layer and training with various encoders, significantly boosted the classification capabilities of the proposed system. Additionally, Alali et al. [45] developed a multi-tasking approach called the Multi-Task Learning Hierarchical Attention Network (MTLHAN), designed to improve sentence representation and increase adaptability. The MTLHAN model utilizes a common word encoder and attention network across tasks, applying dual training strategies for analyzing both three-polarity and five-polarity sentiments in Arabic. The results from the experimental evaluations highlight the exceptional effectiveness of this proposed model.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

Table 1. Text classification 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]
ITL	Transformer	HARD, BRAD, LABR	[44]
Multi-Task Learning	Hierarchical Attention over Bi-LSTM	HARD, BRAD, LABR	[45]
MoE Mechanism and ITL	Switch Transformer	HARD, BRAD, LABR	[64]

3. The Proposed Switching Self-Attention text classification Model That Utilizes Reverse Positional Encoding Method for Right-To-Left Text

Transformer models have demonstrated exceptional performance in a wide array of Natural Language Processing (NLP) tasks, including text classification. The standard transformer model [20], known for its multi-head attention mechanism, serves as a fundamental framework for these tasks. Transformer consists of an encoder with numerous layers of multi-head attention (MHA) and feedforward neural (FFN) networks. The multi-head attention (MHA) technique enables the model to evaluate the relevance of words in a sequence based on their semantic relationships, and the FFNs transform the MHA layer's output into a more refined representation. The essence of the transformer lies in its MHA technique, which is based on specific mathematical formulations [47]. This method processes a series of input embeddings, $x_1, \ldots, x_n$, the MHA method derives a collection of contextually attuned embeddings, $h_1, \ldots, h_n$, through the ensuing procedure:

$$h_i = Attention\ (QW_i^Q, KW_i^k, VW_i^V) \tag{1}$$

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:

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

Additionally, position-wise feedforward networks (FFNs) are essentially multi-layer perceptron that function independently at each position in the sequence. These FFNs provide a non-linear transformation of the outputs from the attention mechanism. The computation of these networks adheres to the following sequence:

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

Within each layer, layer normalization is implemented to standardize the inputs in a neural network, thereby improving the training's efficiency and consistency.

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

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

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 transformer architecture's operational procedure can be effectively outlined in the following manner: Firstly, the input sequence undergoes a linear transformation, producing three distinct vectors - query Q, key K, and value V - achieved by linearly transforming the embedded input. These vectors are then segmented into multiple units known as heads h_i, allowing the model to simultaneously process different aspects of the input sequence, as detailed in Equation (1). Each head h_i applies scaled dot-product attention, determining the attention weights between Q and K by scaling their dot products in proportion to the square root of the vector's dimension. This step assesses the relevance of each K vector to its Q counterpart. Following this, a SoftMax function normalizes the attention weights, ensuring their total equals 1. These normalized weights are then used to modulate the V vectors, creating an attention output for each h_i, as demonstrated in Equation (2). The attention outputs from all the heads are consolidated and re-mapped back to the original vector dimension through an additional linear transformation, as shown in Equation (3). Finally, the composite output is passed through a feedforward network. This introduces nonlinearity into the system, enabling the model to discern more complex relationships between the inputs and outputs, as outlined in Equation (4).

Implementing these methods in every layer of the encoder, the Multi-Head Attention (MHA) mechanism enhances the transformer's ability to discern complex semantic relationships between words in a sequence, making it highly effective for a range of text classification task for ADs. However, the standard transformer architecture faces certain challenges. A significant issue is the MHA mechanism's increasing computational demands in proportion to the length of the input sequence. This poses scalability challenges for very long sequences [48] and limits its flexibility in handling shorter sequences. Moreover, the MHA approach processes all positions in the input sequence uniformly, which may not be ideal for cases where some positions are more critical than others. Despite the transformer's impressive capabilities in various NLP applications, it may encounter difficulties in addressing complex input-output relations that require more specialized modeling approaches.

In response to these challenges, our study presents an innovative Switching Self-Attention text classification (SSA-TC-RPE) model, integrating Reverse Positional Encoding (RPE) and Inductive Transfer Learning (ITL) for the analysis of Arabic dialect (AD) sentiments. The application of inductive transfer learning (ITL) in this model aims to improve the efficiency of five-point Arabic text classification by leveraging the connections between different AD SA classification tasks, including both five-point and ternary polarities. Our SSA-TC-RPE model for ADs builds upon the transformer architecture as initially introduced by Vaswani et al. [20]. ITL has shown to be more effective than single-task learning, as it utilizes a shared representation of various loss functions. This approach simultaneously manages SA tasks involving three and five polarities, thus enhancing the representation of both semantic and syntactic elements in AD texts. Additionally, the knowledge gained from each task reinforces the learning in other tasks, thereby increasing their overall performance. Furthermore, Additionally, a critical aspect of Inductive Transfer Learning (ITL) is its advanced capability to utilize resources designed for similar tasks, thereby enhancing the learning effectiveness of the current task and expanding the pool of usable knowledge. Through this understanding, the layers engaged in sharing tasks can boost the model's ability to generalize, quicken the learning process, and improve its overall clarity. In the same vein, by capitalizing on the domain-specific knowledge inherent in the training signals of related tasks as an inductive bias, the multi-task learning strategy enables efficient transfers that strengthen the model's generalization capacity. Inductive transfer can be utilized to sharpen the precision of generalization, speed up the learning trajectory, and increase the clarity of the developed models. A learner that simultaneously handles multiple related tasks can use these tasks to mutually inform and guide each other, leading to a deeper grasp of the domain's patterns. This approach can be particularly effective in the text classification (TC) of Arabic dialects (ADs), even when training data is limited. In the same way, inductive transfer learning effectively identifies the significant connections among various tasks. As shown in Figure 1, the proposed SSA-TC-RPE text classification framework features a unique design

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

that includes multi-head attention (MHA), RPE technique, ITL, a shared vocabulary, and innovative elements like the mixture of experts (MoE) within the switching FNN layer. The SSA-TC-RPE model, using ITL, optimizes combined classification tasks (ternary and five-polarity) and understands them in unison. The incorporation of a shared Switching Self-Attention block (encoding layer) facilitates the transfer of insights from the ternary task to the five-point task during learning, thereby enhancing the learning effectiveness of the current task (five-point task).

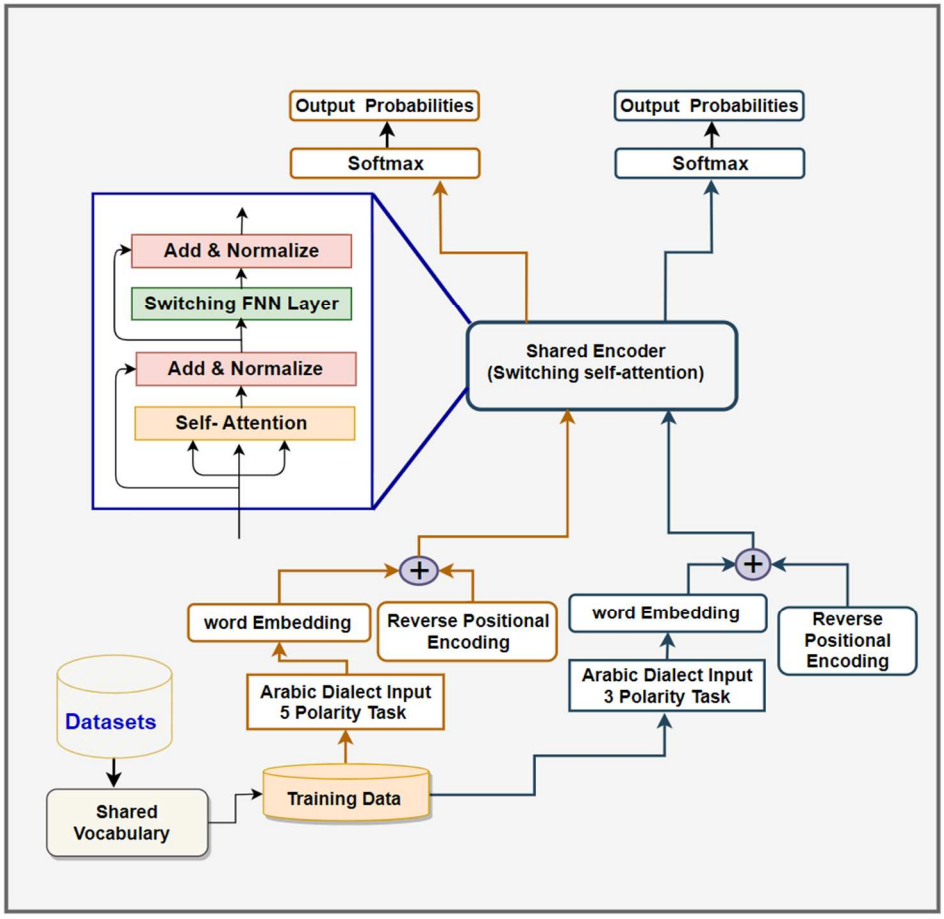


Figure 1. The Architecture of the proposed Switching Self-Attention Text Classification model for Right-to Left text that utilizes the Reverse Positional Encoding (RPE) Method.

Switching self-attention encoder (SSTs) [19] aim to overcome the constraints of classic transformer models by incorporating a sophisticated mixture of experts (MoE) mechanism and RPE method. MoE approach simplifies intricate problems into smaller, more manageable segments, thereby increasing the model's efficiency in handling extensive sequences and intricate input-output correlations. While the multi-head self-attention feature of conventional transformers excels in detecting semantic links within longer sequences, it struggles with shorter ones. The MoE system addresses this by breaking down sequences into smaller sections and assigning them to specific expert networks, thereby enhancing performance and precision for tasks involving shorter sequences, as demonstrated by improved results in various benchmark assessments [49–51]. A key development in Switching self-attention encoder is substituting the typical feed-forward network (FFN) with the MoE mechanism, as illustrated in Figure 2. Standard transformers employ an FFN comprising two linear layers and a ReLU activation function, whereas the MoE method uses a collection of expert networks that dissect and analyze distinct features of the input, combining their findings through a gating network. This enables the model to dynamically choose from an array of parameters or expert modules based on the specific input, deviating from the traditional transformer's static parameter usage, as explained in Equation 4. Formally, the MoE mechanism in Switching self-attention encoder is represented by the following equation:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

$$z_t = \sum_j g_j(x_t) * e_j(x_t) \quad (6)$$

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 functioning of the switch mechanism involves training the parameters of the gating functions, which facilitates the dynamic selection of different expert modules. This adaptability allows the model to efficiently handle various input patterns and perform effectively across multiple tasks. The core operation of the Mixture of Experts (MoE) mechanism in the switching self-attention involves several key steps. Initially, the input is divided into several segments, with each segment being independently processed by a specific expert. These experts are individual neural networks, each specializing in a particular segment of the input data. They produce an output vector representing their prediction for the assigned input segment. A gating mechanism then determines the most relevant expert for a given input. This mechanism evaluates the input and calculates a set of weights that determine the importance of each expert's prediction. The final output of the model is a composite of these expert predictions, with the weighting of each prediction governed by the gating process. For inductive transfer learning (ITL), the loss and optimizer are engaged sequentially for each task. The training alternates between spending a set number of cycles on the ternary classification task and then focusing on the five-polarity classification tasks. The goal in training both tasks concurrently is to reduce cross-entropy, leading to the achievement of more effective learning outcomes.

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

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

where $\hat{y}_j^i$ and y_i^j 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 SSA-SA system, we received the following global loss function:

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

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 ($\lambda = 1$). Overall, the Mixture of Experts (MoE) endows the switching self-attention encoder with the ability to grasp complex patterns in the input data by leveraging the unique skills of various expert networks. This setup allows the model to integrate insights from multiple experts, each proficient in different aspects of the data, and merge their contributions for improved performance. Such an approach is particularly effective in tasks requiring deep understanding of inputs, offering a promising solution to the challenges posed by limited datasets in Arabic Dialect (ADs) text classifications. Thus, this research utilizes this capability to decode complex interactions between words and phrases in ADs texts.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

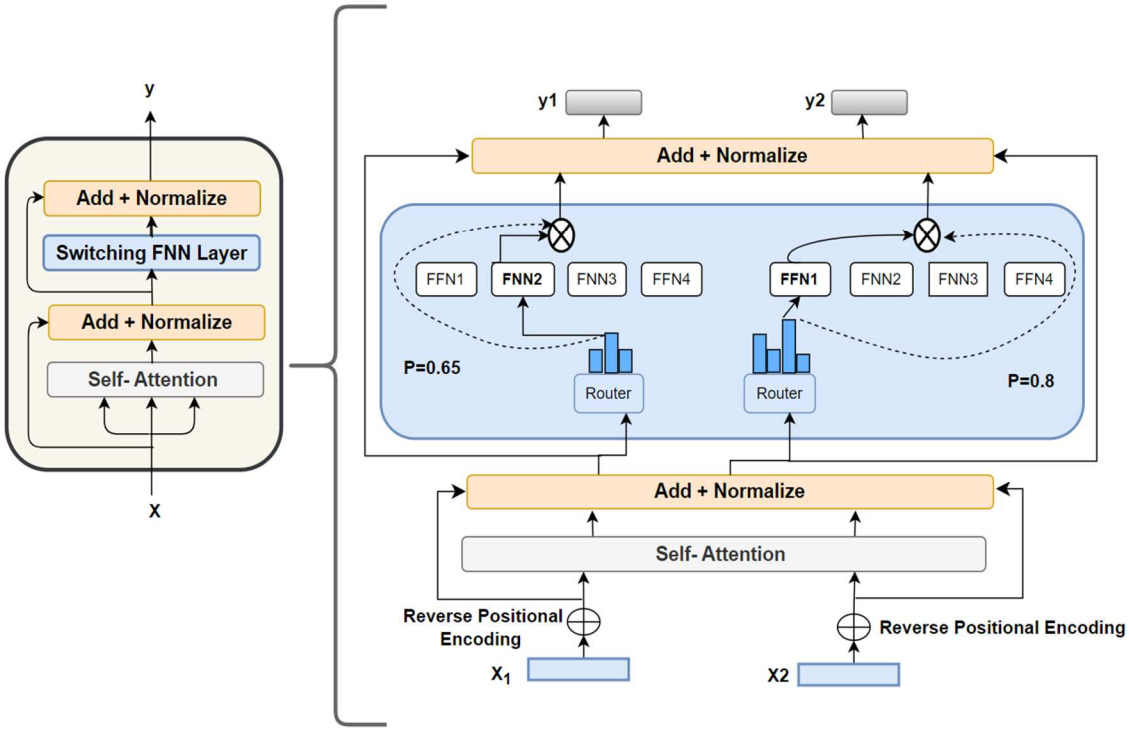


Figure 2. Detailed architecture of the encoder and switching FNN layer in the proposed SSA-TC-RPE model.

3.1. Reverse Positional Encoding (RPE)

Word arrangement and sequence are crucial components in the structure of any language. Changing the order of words can significantly alter the meaning of a sentence. For natural language processing applications, the inherent functionality of Recurrent Neural Networks (RNNs) effectively manages the sequencing of words. This involves parsing the sentence in a linear manner, analyzing each word one after the other. Such a method merges the sequential order of words with the foundational framework of RNNs. In contrast, the transformer model eschews the use of RNNs, favoring a self-attention mechanism that processes each data point on an individual basis. Eliminating the RNN layers can speed up the training process and enhance the model's ability to grasp long-range dependencies within a sentence. In the transformer architecture, as words from a sentence are processed simultaneously through its encoder/decoder, the model initially lacks awareness of the words' sequence or position. Therefore, it's necessary to incorporate a system that embeds the sequence of words into the model. A practical approach to achieve this is by appending additional data to each word, indicating its position in the sentence. This additional data is referred to as "reverse positional encoding" (RPE). RPE serves as a technique to retain information about the order of elements in a sequence.

Reverse Positional Encoding (RPE) assigns a distinct representation to each position in a sequence, thereby defining the placement of an entity within it. In our newly developed proposed model, we implement an innovative method where reverse positional encoding is utilized to capture the sequential relationships among words in an AD sentence, as illustrated in Figure 3. In texts written in Arabic dialects, which are read from right to left, the sentence begins with the first word on the right and concludes with the last word on the left. Considering the embedded sequence of the ADs input sentence, which has a length of J, denoted $X = \{x_1, \ldots\ldots, x_J\}$, where x_1 corresponds to the embedding of the final word in the AD sentence, and x_J corresponds to the initial word in the ADs input sentence. The reverse positional embedding for each word is computed based on its position, as defined in Equation (11).

$$Rpe\ (j, 2i) = sin\left(j/10000^{2i/d_{model}}\right),$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

$$Rpe\ (j, 2i+1) = cos\left(j/10000^{2i/d_{model}}\right) \quad , \tag{11}$$

j represents the location of a word within a sentence, while i denotes the number of dimensions corresponding to that positional index. Consequently, a sequence of reverse positional embeddings is established

$$RPE = \{Rpe_j, \dots, Rpe_1\}. \tag{12}$$

Each $\boldsymbol{Rpe_j}$ is summed to the corresponding word embedding $\boldsymbol{x_j}$ as united embedding $\boldsymbol{v_j}$:

$$v_j = x_j + \boldsymbol{Rpe}_j \tag{13}$$

Finally, a sequence of embeddings $\{v_1, \dots, v_J\}$ is sentence representation that will be initialized H^0. Subsequently, H^0 is inputted into the self-attention sublayer for sentence representation analysis. Similar to how humans classify sentences by modifying word orders based on the original sentence's semantic and contextual aspects to produce an easily understandable synonymous sentence, the proposed model applies reverse positional encoding. This encoding, which takes into account the details of the input sentence, is specifically integrated into the encoder. A comprehensive explanation of the RPE method is clarified in Figure 4.

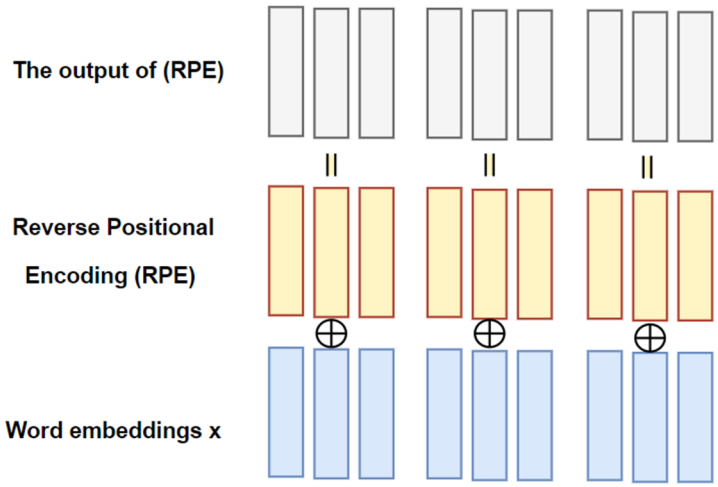
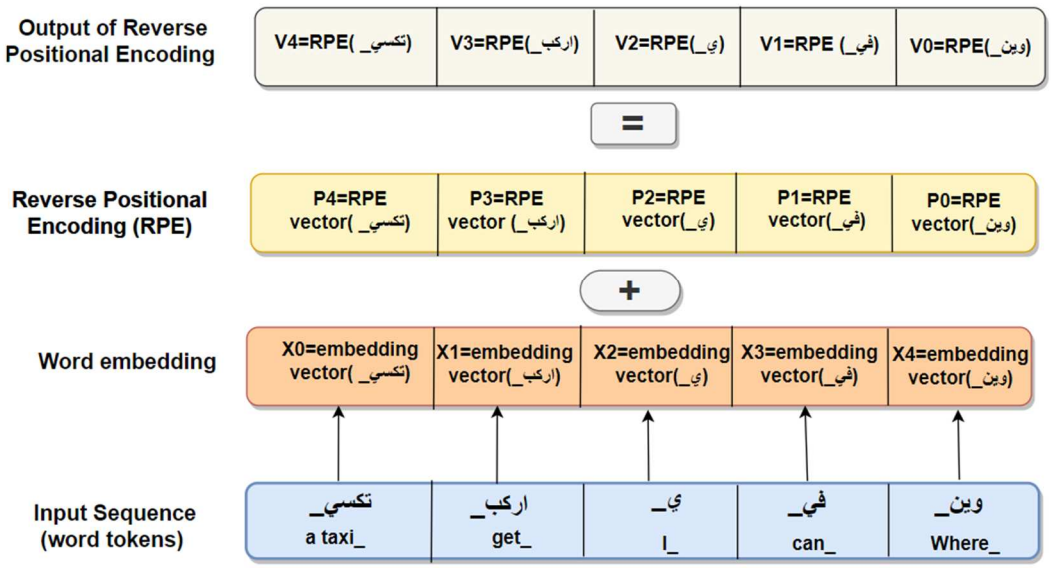


Figure 3. Reverse Positional Encoding Mechanism



Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

Figure 4. A detailed structure of the Reverse Positional Encoding Mechanism.

4. Experiments

A range of empirical assessments was carried out to determine the efficacy of the SSA-TC-RPE model in classifying Arabic vernaculars. The capability of the proposed SSA-TC-RPE model in categorizing Arabic dialects (ADs) was comprehensively scrutinized.

4.1. Data

The Training for the proposed model was executed using three key datasets. The primary dataset was HARD [27], consisting of reviews from a variety of booking sites, sorted into five distinct classes. This was followed by employing the BRAD [26] and LABR [22] datasets for further training. This study made use of datasets at the review level, incorporating BRAD, HARD, and LABR. Specifically, BRAD's reviews, sourced from the Goodreads website, were classified into five distinct scales. The class distributions for HARD, BRAD, and LABR are outlined in Tables 2–4, respectively. It should be noted that the datasets in this study were used in their original, raw form, which might have implications for the accuracy of the model. Furthermore, preprocessing was performed on all sentences, involving sentence segmentation to break down the reviews into discrete sentences. This procedure also included the elimination of Latin characters, non-Arabic elements, diacritics, hashtags, punctuation, and URLs from the ADs texts. The Arabic dialect texts underwent orthographic normalization to ensure consistency and standardization [2]. Emoticons within the data were converted into corresponding textual descriptions, and adjustments were made for words that were artificially lengthened. To prevent the risk of the model becoming over-fitted, an early stopping mechanism was implemented, setting the patience parameter at three epochs. For assessing the SSA-TC-RPE model's effectiveness, which incorporates Inductive Transfer Learning (ITL) in classifying Arabic dialect texts, a checkpoint system was utilized to save the most optimal weights of the model. The dataset was divided, with 80% allocated for training and 20% for testing. Additionally, a K-fold cross-validation technique with k = 2 was employed to establish a split between training and testing for the model's evaluation [52]. Examining the HARD, BRAD, and LABR datasets revealed the sentiment distribution within these samples. The HARD dataset contained 409,562 entries, categorized into 5 sentiment types. Allocating 80% of this dataset (327,649 samples) for training and the remaining 20% (81,912 samples) for testing allowed for a comprehensive understanding of sentiment variation. In a similar manner, the BRAD dataset with 510,598 entries was divided, with 80% (408,478 samples) used for training and 20% (101,019 samples) for testing. The smaller LABR dataset, consisting of 63,257 entries, also maintained the same 80–20 division for training (50,606 samples) and testing (12,651 samples). These splits ensured that all five sentiment categories were well-represented in both training and testing stages, aiding the models in grasping sentiment subtleties and applying this understanding to new data. Biases in text classification models can significantly affect their accuracy. If training data contain biases, they might distort the results. To mitigate this issue and ascertain the optimal data selection for our SSA-TC-RPE text classification model tailored for Arabic dialects, we executed five specific steps:

- Ensured that the training dataset included a diverse range of sources, covering various demographic groups, geographic areas, and social environments. This method was aimed at reducing biases, leading to a dataset that was not just comprehensive but also balanced in its representation.
- Verified that the sentiment labels within the training dataset were uniformly allocated across all demographic groups and perspectives
- Set Established clear guidelines for labeling, instructing human annotators to maintain neutrality and avoid infusing their own biases into the sentiment labels. This strategy helped ensure consistency and minimize the likelihood of bias in the dataset.
- Undertook a thorough review of the training data to identify any underlying biases. This involved examining aspects such as demographic imbalances, stereotype reinforcement, and potentially underrepresented groups. Once these biases were detected, corrective actions were taken. These included using methods like data augmentation, increasing the representation of underrepresented groups through oversampling, and implementing various preprocessing techniques.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

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 newly proposed SSA-TC-RPE model, termed the switching self-attention and employing RPE technique, MoE mechanism and inductive-transfer learning (ITL), was built utilizing the features of the TensorFlow [53], Keras [54], and scikit-learn [55] frameworks. To evaluate its performance, extensive experiments were conducted on classification tasks for Arabic Dialects (ADs), covering both three and five polarity levels. These tests included a wide range of parameter settings, particularly testing six different word-embedding dimensions for each token, namely 50, 32, 40, and 35. The model's attention heads were also varied, experimenting with six specific values: 4, 2, and 3. Furthermore, the position-wise switching-Feedforward Neural Network (FNN) was tested with filters of different sizes, including 40, 30, 35, 32, and 50.

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

In inductive-transfer learning models, joint training and alternative training stand as two primary methodologies. Joint training simultaneously trains a model across multiple tasks, enabling shared learning and representation beneficial to all tasks, and capitalizes on task interdependencies to enhance overall effectiveness. On the other hand, alternative training sequentially focuses on individual tasks, allowing the model to dedicate its resources to each task in turn. This can lead to improved performance in each distinct task. Both strategies offer their own advantages and challenges. Joint training tends to promote better generalization across various tasks, while alternative training may be more effective for tasks with significant differences in data distribution or complexity. The choice between these strategies hinges on the specific characteristics of the tasks at hand and the desired balance between performance and efficiency. Ultimately, selecting the appropriate training method is crucial for optimizing the success and adaptability of inductive-transfer learning models.

The developed system adeptly handled both ternary and five-tier sentiment classification tasks. For instance, with the HARD dataset, the SSA-TC-RPE model alternated between processing five-tier and ternary classification tasks. We utilized two distinct training methods: an alternating [1,2] strategy and a simultaneous joint learning technique. In our inductive transfer learning framework, the loss function and optimizer were applied in sequence to each task. The training commenced with the ternary classification task for a predetermined number of epochs before transitioning to the five-polarity task, aiming to minimize categorical cross-entropy for both. Training of the SSA-TC-RPE model spanned 20 epochs, featuring an early stopping mechanism activated after two epochs without progress, with a batch size of 90. We followed standard protocols for the BRAD, HARD, and LABR

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

datasets, allocating 80% for training and 20% for testing. The Adam optimizer was employed for each task in the SSA-TC-RPE model. Sentence segmentation was used to break down reviews into individual sentences, with maximum lengths set to 80 for BRAD, 50 for HARD, and 80 for LABR. Our model did not incorporate class weights [56]. To facilitate effective learning, training data were shuffled at the start of each epoch. More details on the hyper-parameters and their configurations are provided in Section 4.5.

4.4. Sate-of-the-Art Approaches

Using the five-point datasets BRAD, HARD, and LABR for Arabic dialects (ADs) analysis, we evaluated the specifically designed SSA-TC-RPE model against contemporary benchmark methods. Initially, logistic regression (LR) using unigrams, bi-grams, and TF-IDF was proposed in [26] for the BRAD dataset. Similarly, the LR method with analogous features was advocated in [27] for the HARD dataset. The SSA-TC-RPE model was also subjected to a comparative analysis using the LABR datasets. This comparison involved benchmark methods like SVM, employing a support vector machine classifier with n-gram features as indicated in [23], MNB using a multinomial naive Bayes approach with bag-of-words features as detailed in [22], and HC, a hierarchical classifiers model based on the divide-and-conquer strategy referenced in [24]. The HC(KNN) model, an enhanced hierarchical classifier variant, continues to rely on the divide-and-conquer method as outlined in [25]. In recent developments in natural language processing (NLP), significant progress has been achieved with the bidirectional encoder representations from transformers, or BERT [57]. In particular, AraBERT [58], an Arabic-pre-trained BERT model, underwent training on diverse datasets including OSIAN [59], Arabic Wikipedia, and the MSA corpus, totaling about 1.5 billion words. We conducted a comparative analysis of our SSA-TC-RPE model for Arabic dialects against other models such as AraBERT [58] and T-TC-INT [44], ST-SA [64] assessing their efficacy in different scenarios.

4.5. Results

A series of experimental tests were conducted using the proposed SSA-TC-RPE model for Arabic dialects. We trained the SSA-TC-RPE system with various configurations, including different numbers of attention heads (AHs) in the Multi-Head Attention (MHA) sub-layer and varying quantities of encoders to determine the most effective structure. The system also underwent training with different word embedding sizes for each token. This study evaluated the impact of employing two inductive transfer learning approaches – concurrent and alternating – on the system's performance. The effectiveness of the SSA-TC-RPE system in text classification was measured using an automated accuracy metric. This part of the research presents an evaluation of the SSA-TC-RPE system's performance in five-polarity classification tasks for Arabic dialects. The outcomes of these empirical tests on the HARD, BRAD, and LABR datasets are presented in Tables 5–7. As detailed in Tables 5 and 8, the SSA-TC-RPE system demonstrated high performance on the HARD imbalanced dataset, achieving an accuracy of 87.20%, an F-score of 86.91%, and a precision of 86.50%. This was accomplished with a configuration of 2 attention heads (AH), 80 tokens, 10 experts, a batch size of 65, a filter size of 30, a dropout rate of 0.20, and an embedding dimension of 25 for each token. The system's notable accuracy is attributed to the effective integration of the Inductive Transfer Learning (ITL) framework, Mixture of Experts (MoE) mechanism, and Multi-Head Attention (MHA) approach, especially for right-to-left scripts like Arabic Dialects (ADs). The MoE mechanism uses a series of expert networks that analyze different aspects of the input data and combine their outputs through a gating network. This feature allows the model to dynamically choose from a range of parameters (expert modules) based on the input, enhancing its ability to accurately discern sentiments. Comparing the SSA-TC-RPE model's results to the top-performing system on the HARD dataset, it outperformed the Logistic Regression (LR) [60] model by 11.1% in accuracy. Moreover, the developed model excelled beyond AraBERT [58], achieving an accuracy that was 6.35% higher, and also outdid the T-TC-INT model [45] by a margin of 5.37% in accuracy. Also, the proposed model outperformed the ST-SA model [64] by 3.18% in accuracy. This superior performance can be attributed to the simultaneous processing of related learning tasks, which broadened the data spectrum and

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

minimized the likelihood of overfitting [61]. The system demonstrated adeptness in recognizing both syntactic and semantic elements, enabling precise detection of sentiments expressed in Arabic Dialect (AD) sentences.

Table 5. Results for the proposed SSA-TC-RPE 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
60	40	4	50	15	50	0.20	84.49%	84.00%	83.97%
35	90	2	30	10	55	0.25	85.17%	84.30%	84.10%
25	80	2	30	10	65	0.20	**87.20%**	86.91%	86.50%
50	100	4	30	5	60	0.25	84.76%	84.18%	84.05%
30	30	4	35	5	55	0.30	83.87%	83.14%	83.30%

Table 6. Results for the SSA-TC-RPE 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
23	30	4	30	10	50	0.30	70.85%	70.09%	70.10%
45	20	3	30	12	60	0.25	68.75%	67.50%	67.81%
40	25	4	35	15	45	0.30	69.97%	69.74%	69.50%
55	22	4	35	18	50	0.25	**72.17%**	71.89%	71.72%
60	25	4	30	20	65	0.30	70.11%	70.02%	69.92%

Table 7. Results for the SSA-TC-RPE 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
45	25	4	30	10	45	0.30	84.50%	83.20%	83.40%
65	90	2	30	11	80	0.20	**86.89%**	85.91%	85.39%
40	35	2	30	12	50	0.25	85.48%	85.00%	84.70%
30	25	4	40	13	50	0.20	84.90%	84.30%	83.95%
60	45	2	35	10	40	0.20	85.08%	84.28%	84.15%

Table 8. The performance of the proposed SSA-TC-RPE 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%
SA-SA Model [64]	5	84.02%	83.50%
Proposed SSA-TC-RPE	5	87.20%	86.91%

Moreover, the suggested SSA-TC-RPE system demonstrated remarkable results on the imbalanced BRAD dataset. As depicted in Table 6, the model attained an accuracy of 72.17%, an F-score of 71.89%, and a precision of 71.72%. These outcomes were achieved with configurations of 4 attention heads (AH), 22 tokens, 18 experts, a batch size of 50, a filter size of 35, a dropout rate of 0.25, and an embedding dimension of 55 for each token. As indicated in Table 9, the SSA-TC-RPE system significantly outperformed the logistic regression (LR) method [26], showing an accuracy

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

improvement of 24.47%. It also exceeded the AraBERT model [58] by 11.32% and the T-TC-INT [44] system by 10.44%. It also exceeded the SA-ST model [64] by 3.36%. Fu rthermore, the use of a switching self-attention based shared encoder, one for each classification task, enabled the model to effectively represent the context before, after, and around any position within a sentence, leading to a more nuanced and comprehensive understanding.

Table 9. The performance of the proposed SSA-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%
SA-SA Model [64]	5	68.81%	67.89%
Proposed SSA-TC-RPE	5	72.17%	71.89%

Additionally, as outlined in Table 7, the proposed Switching Self-Attention text classification model utilizing multi-task learning (SSA-TC-RPE) exhibited remarkable results on the complex and imbalanced LABR dataset. In this research, the model impressively achieved an accuracy rate of 86.89%, an F-score of 85.91%, and a precision of 85.39%, outperforming other methods. It's notable that this high level of performance was attained with specific configurations, including two attention heads (AHs), a filter size of 30, 90 tokens, 11 experts, a batch size of 80, a dropout rate of 0.20, and an embedding dimension of 65 for each token. These results underscore the efficacy of the SSA-TC-RPE model in tackling the intricacies of text classification in an imbalanced dataset, demonstrating its robust capabilities. Table 10 showcase the superior performance of the proposed Switching Self-Attention text classification model employing inductive transfer learning (SSA-TC-REP) when compared to a range of alternative methodologies. The SSA-TC-RPE model notably exceeded several other models by considerable margins. For instance, it surpassed the SVM [23] model with an impressive accuracy increase of 36.59%, outdid the MNP [22] model by 41.89% in accuracy, exceeded the HC(KNN) [24] model by 29.09%, and achieved an accuracy that was 27.93% higher than AraBERT [58]. Furthermore, it outstripped the HC(KNN) [25] model by 14.25% in accuracy. The model also performed better than the T-TC-INT model [44], showing an accuracy improvement of 8.76%. Furthermore, the proposed model also performed better than the SSA-TC-RPE model [64], showing an accuracy improvement of 2.98%.

Table 10. The performance of the proposed SSA-TC-RPE 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%
SA-SA Model [64]	5	83.91%	82.71%
Proposed SSA-TC-RPE	5	86.89%	85.91%

Within the realm of deep learning, joint training involves training a single neural network on several interconnected tasks at the same time. Rather than developing individual models for each task, this strategy enables the model to recognize and utilize shared characteristics across all tasks, thereby enhancing its versatility and operational efficiency. This method typically leads to improved outcomes for each task, as the model benefits from the synergistic relationships between the tasks. In contrast, imbalanced data refers to datasets in which the distribution of classes (or categories) is

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

uneven. This often means that one or more classes are underrepresented compared to others, presenting potential difficulties in training the model and assessing its performance.

This scenario poses challenges for deep learning models, as they may develop a tendency to favor the more prevalent class, leading to inferior performance on the less represented classes. The evaluation results indicate that the SSA-TC-RPE system, when applied to both joint and alternative learning methods, shows remarkable effectiveness. Alternative training outperformed joint training, as evidenced by higher accuracies of 87.20% and 81.79% on the imbalanced HARD dataset, and 72.17% and 68.21% on BRAD, respectively, as shown in Table 11. In contrast to conventional approaches, alternative training in a five-tier classification model seems to better capture subtle feature variations in text sequences compared to single-task learning. These results suggest that alternative learning is more suitable for complex text classification (TC) tasks, enabling the development of more intricate and detailed latent representations for Arabic Dialect text classification (ADs TC) tasks. The distinct performance difference between the two methodologies can be ascribed to how alternative training utilizes the diverse data volumes available in the datasets of each task.

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

SSA-SA Training Method	HARD (Imbalance) Accuracy	BRAD (Imbalance) Accuracy
Alternately	87.20%	72.17%
Jointly	81.79%	68.21%

Shared layers often contain a greater amount of information for tasks with larger datasets. However, joint learning may exhibit a bias towards tasks associated with significantly larger datasets. Therefore, alternative training methods are generally considered more appropriate for tasks like text classification of Arabic dialects. This is especially the case when dealing with two separate datasets for different tasks, such as in machine translation scenarios where the transition is from Arabic dialects (ADs) to Modern Standard Arabic (MSA) and then to English [2]. By alternating the network's focus between tasks, the efficiency of each task is enhanced without the need for additional training data [1]. Additionally, harnessing the synergistic relationship between related tasks can improve the efficacy of five-point classification systems. The marked improvements in our model's performance can be ascribed to several factors. Outperforming established models like AraBERT, renowned for its proficiency in Arabic language tasks, is a notable accomplishment. Our model's ability to exceed AraBERT's performance on the same datasets demonstrates its enhanced precision in processing Arabic dialects. Even small improvements in accuracy are valuable, as they contribute to the overall development of models tailored for Arabic dialect processing. These advancements can have practical applications, such as in more accurate text classification, better information retrieval, and other natural language processing tasks designed for Arabic dialects.

The SSA-TC-RPE system notably did not show significant improvements on the BRAD dataset when compared to existing models. This lack of enhanced performance might stem from the system's limited understanding of the distinct characteristics, idiomatic expressions, and linguistic subtleties unique to the BRAD Arabic dataset. A lack of adequate domain-specific adaptation could lead to a disconnect between the features the model learns and the unique elements of the BRAD dataset, resulting in suboptimal performance. To boost the model's effectiveness in text classification on the BRAD dataset, implementing advanced deep learning techniques, particularly domain adaptation, is essential. For instance, the use of transformers, especially BERT (Bidirectional Encoder Representations from Transformers), has been transformative in NLP due to their proficiency in contextualizing text. Optimizing a pre-trained BERT model specifically for the BRAD dataset could substantially improve its text classification capabilities.

In evaluating the practicality, while pre-trained models are easily accessible, fine-tuning them demands significant computational power and NLP expertise. This process is viable with the availability of these resources. Adversarial training is another crucial method, where the model is trained to withstand manipulative adversarial examples. For text classification involving five-polarity Arabic dialects, this approach can enhance the model's ability to deal with subtle and varied expressions of sentiment. Although the implementation of adversarial training can be intricate and

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

resource-intensive, it is achievable with sufficient deep learning resources and know-how. Domain-adaptive fine-tuning stands out as an effective technique, particularly for text classification on the BRAD dataset. It involves incrementally fine-tuning a pre-trained model using a combination of data from both the source and target domains, with an increasing emphasis on the target domain. This method facilitates the model's adaptation to the unique linguistic features and sentiment expressions specific to the BRAD dataset.

Moreover, domain-adaptive fine-tuning is a viable option when there is an ample supply of data from both the source and target domains. It requires fewer resources than building a model from the ground up. Meta-learning, another approach, trains a model across a variety of tasks, enabling it to quickly adapt to new tasks or domains. This method proves beneficial in five-polarity text classification for Arabic dialects (ADs), as it can accommodate a wide range of expressions and contexts. However, meta-learning needs diverse training datasets and substantial computational resources, making it suitable for environments with ample resources. In cases where the BRAD dataset encompasses multilingual data, cross-lingual models like multilingual BERT become effective. These models, trained in several languages, are adept at conducting text classification across different linguistic scenarios. Pre-trained versions of these models, akin to BERT, are available. Fine-tuning them for the specific languages in the BRAD dataset is essential and can be done with the right computational infrastructure.

4.6. Impact of Number of Experts (NE)

As illustrated in Tables 5–7, the efficacy of the proposed SSA-TC-RPE model in processing various input types from the self-attention layer highlights its importance for the classification task involving five distinct sentiment polarities. In this context, "NE" represents the count of experts in the encoding layer of the Switching Self-Attention SA model, which utilizes the RPE mechanism. The model was trained with different numbers of experts, including 5, 10, 11, 12, 13, 15, 18 and 20. The data presented in Tables 5–7 show a noticeable variation in accuracy scores across the HARD, BRAD, and LABR categories, depending on the number of experts used.

4.7. Impact of Length of Input Sentence

Gaining a deeper understanding of extended syntactic dependencies and contextual connections among elements in input phrases significantly improves the ability to classify longer sentences. Following the methodology of Luong et al. [62], sentences with similar lengths (measured by the number of source tokens) were grouped together. Given the considerable size of the HARD corpus, a task focusing on a five-category classification of sentiments within the HARD dataset was chosen to evaluate the effectiveness of the self-attention (SA) mechanism in analyzing long sentences. The assessment in this section was predicated on the subsequent ranges: <10, 10–20, 20–30, 30–40, 40–50, and >50.

The performance of the Switching Self-Attention text classification system was evaluated using an automated accuracy metric. As shown in Table 12, the effectiveness of the suggested Switching Self-Attention text classification (SSA-TC-RPE) model improved with the increase in sentence length. This enhancement was particularly notable in sentences containing 40 to 50 word tokens, as well as those with more than 50 word tokens, achieving accuracy rates of 87.20% and 85.87%, respectively. The system's success in capturing contextually relevant knowledge and dependencies of the tokens, irrespective of their placement in the Arabic Dialect (AD) input sentences, was attributed to the utilization of ITL, RPE technique, switching FNN sub-layer, and the integration of word units as an input feature for the Multi-Head Attention (MHA) sub-layer. The adoption of the REP in the switching self-attention encoder significantly enhanced its ability to identify complex patterns in the input data, utilizing the expertise of numerous specialists. Nevertheless, the performance of the proposed model was relatively lower with shorter sentences, particularly those consisting of 10 to 20, 20 to 30, and 30 to 40 word tokens. Additionally, the system's efficiency markedly decreased for sentences under 10 words, recording a modest accuracy of 76.46%. The commendable performance of the SSA-TC-RPE system across a range of sentence lengths highlights the effectiveness of combining the RPE technique and ITL framework. This combination, along with the integration of

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

the MoE mechanism, significantly improved the capability of the encoder's MHA sublayer in analyzing word relationships within Arabic Dialect (AD) input sentences.

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

Sentence Length	Accuracy
<10	76.46%
(10–20)	78.14%
(20–30)	80.04%
(30–40)	82.63%
(40–50)	85.87%
>50	87.20%

4.8. Motivation and Novelty

Regarding text classification in Arabic Dialects (ADs), our study presents a novel Switching Self-Attention text classification (SSA-TC-RPE) model, specifically designed for the five-level classification of ADs. This initiative was driven by the necessity for efficient text classification techniques in a field marked by complex linguistic features and scarce training data. The key innovations and driving factors of this research are summarized as follows:

1. Improving textual representation through Inductive Transfer Learning (ITL): Our method involved integrating an Inductive Transfer Learning (ITL) framework with the self-attention mechanism. This innovative blend was designed to enhance the representation of text sequences, aiming to bolster the system's capacity to assimilate both comprehensive and detailed semantic insights within the given context.
2. Our study focused on the issue of imbalanced data text classification, especially within the framework of Arabic Dialects (ADs). This consideration was vital for ensuring precise sentiment classification in an area where some sentiments might be underrepresented.
3. Employing the Reverse Positional Encoding (RPE) technique in our model facilitated the detection of crucial terms and words in text sequences. This approach notably enhanced the model's ability to grasp the subtleties inherent in Arabic Dialects (ADs).
4. In-depth text classification: Our research delved into a nuanced text classification method, incorporating ternary classifications in the context of Inductive Transfer Learning (INT). This strategy was aimed at improving the precision of sentiment differentiation, particularly distinguishing more accurately between high-negative and negative sentiments within the framework of the five-level classification system.
5. Outstanding Results: The empirical findings showcase the exceptional performance of our SSA-SA model compared to current leading-edge approaches, evident across various datasets like HARD, BRAD, and LABR. These results underscore the practical efficiency of our methodology in real-world scenarios.
6. Managing Syntactic Challenges: Our model successfully navigated the constraints of limited training data and skillfully tackled the inherent syntactic intricacies of the unstructured format typical of Arabic Dialect (AD) phrases. This distinctive proficiency distinguishes our SSA-SA system.
7. Integrating Progressive Methods: The SSA-TC-RPE system employed state-of-the-art methodologies, such as the Reverse Positional Encoding (RPE) technique, Multi-Head Attention (MHA) strategy, Mixture of Experts (MoE) mechanism, and the use of word units as input features. These combined advancements led to the development of an exceptionally effective sentence classification system, specifically designed for Arabic Dialects (ADs).

5. Conclusions

We developed an SSA-SA model specifically for the five-level classification of Arabic dialects (ADs). This innovative model leverages a self-attention mechanism and integrates a inductive transfer learning (ITL) framework to improve text sequence representation. The Multi-Head

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

Attention (MHA) method in this model effectively identifies key terms and words in text sequences. Training the model on text classification (TC) tasks, which include both ternary and five-polarity categories for ADs, significantly boosted its efficiency. The combination of MHA with ITL significantly improved the quality of our SA system. The results from this study highlight the essential features of the SSA-TC-RPE system, which uses the Mixture of Experts (MoE) mechanism and MHA approach to enhance accuracy in both five-point and three-point classification tasks. The incorporation of the ITL framework, MoE mechanism, and word units as input features in the MHA sub-layer illustrates the importance of these techniques in text classification tasks for languages with limited resources, like ADs. Experimenting with a range of setups, such as utilizing multiple heads in the Multi-Head Attention (MHA) sub-layer and training with a variety of expert counts, enabled our SSA-TC-RPE model to adeptly decipher complex patterns in the input by leveraging the distinct skills of numerous experts. This approach significantly enhanced the classification capabilities of our system. Our extensive testing on two datasets for five-level Arabic text classification (TCs) demonstrated that alternative learning approaches were more effective than joint learning, with the size of each task's dataset playing a crucial role. The results clearly indicate that our proposed system outperforms other sophisticated methods when evaluated on the HARD, BRAD, and LABR datasets. Additionally, we found that using alternative training strategies within the inductive transfer learning (ITL) framework markedly boosted the efficacy of the five-level classification. In particular, implementing a comprehensive ternary classification approach, particularly in identifying negative text, contributed to a more precise distinction between high-negative and negative sentiments in the five-level classification scheme.

Experiments on five-point and three-point classification tasks revealed that our recommended system markedly enhanced accuracy over other Arabic dialect text classification models. The developed Switching Self-Attention text classification (SSA-TC-RPE) model, leveraging inductive transfer Learning (ITL), produced robust latent representations for textual sequences in Arabic dialects. Achieving accuracy rates of 87.20% for HARD, 72.17% for BRAD, and 86.89% for LABR, the empirical data highlight the SSA-TC-RPE model's advanced performance compared to contemporary leading approaches, such as AraBERT [58], Support Vector Machine (SVM) [23], Multi-Neural Perceptron (MNP) [22], Hierarchical Clustering with K-Nearest Neighbor (HC(KNN)) [24], Logistic Regression (LR) [26], and T-TC-INT [44], and ST-SA [64]. Further evaluation revealed that the system's success depended on the use of the Multi-Head Attention (MHA) strategy and the specific dimensions of word embeddings for each token. This thorough investigation brought to light the benefits of the MHA method, particularly its ability to extract comprehensive and specific semantic insights within the context, facilitated by the MHA sub-layers in the encoding layers.

The SSA-TC-RPE system, developed for text classification in Arabic Dialects (ADs), effectively addresses both the challenge of limited training data and the syntactic complexities typical in the unstructured format of AD (advertisement) phrases. This system stands out due to its unique approach, which includes state-of-the-art techniques such as the Multi-Head Attention (MHA) strategy, Mixture of Experts (MoE) mechanism, RPE technique, and the inclusion of word units as input for the MHA sub-layer. These combined technologies create an efficient and specialized sentence classification system for ADs, enabling precise text classification within this area. Looking to the future, we plan to further refine the capabilities of the SSA-TC-RPE system. Our current focus is on developing a multi-task learning text classification framework that incorporates sub-word units as inputs for the MHA sub-layer, in line with recent research [52,63]. These enhancements are intended to strengthen the system's proficiency in handling varied linguistic subtleties and to improve the precision of text classification. This would make the SSA-TC-RPE system a more versatile and accurate tool for text classification in ADs across different 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.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

Data Availability Statement: The dataset generated during the current study is available from the [SSA_TC_RPE] repository (https://github.come/laith85, accessed on 1 February 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.
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.
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.
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.
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.
7. Abdul-Mageed, M. Modeling Arabic subjectivity and sentiment in lexical space. *Inf. Process. Manag.* **2019**, *56*, 308–319.
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.
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.
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.
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.
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.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

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.
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.
30. Sohangir, S.; Wang, D.; Pomeranets, A.; Khoshgoftaar, T.M. Big Data: Deep Learning for financial sentiment analysis. *J. Big Data* **2018**, *5*, 3.
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.
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.
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.
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.
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
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.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 February 2024 doi:10.20944/preprints202402.0332.v1

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.
47. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111–132.
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.
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.
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.
64. Baniata, L.H.; Kang, S. Switch-Transformer Sentiment Analysis Model for Arabic Dialects That Utilizes a Mixture of Experts Mechanism. *Mathematics* 2024, *12*, 242.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.