mathematics

MDPI

Article

Transformer Text Classification Model for Arabic Dialects That Utilizes Inductive Transfer

Laith H. Baniata * and Sangwoo Kang *

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

Abstract: In the realm of the five-category classification endeavor, there has been limited exploration of applied techniques for classifying Arabic text. These methods have primarily leaned on single-task learning, incorporating manually crafted features that lack robust sentence representations. Recently, the Transformer paradigm has emerged as a highly promising alternative. However, when these models are trained using single-task learning, they often face challenges in achieving outstanding performance and generating robust latent feature representations, especially when dealing with small datasets. This issue is particularly pronounced in the context of the Arabic dialect, which has a scarcity of available resources. Given these constraints, this study introduces an innovative approach to dissecting sentiment in Arabic text. This approach combines Inductive Transfer (INT) with the Transformer paradigm to augment the adaptability of the model and refine the representation of sentences. By employing self-attention SE-A and feed-forward sub-layers as a shared Transformer encoder for both the five-category and three-category Arabic text classification tasks, this proposed model adeptly discerns sentiment in Arabic dialect sentences. The empirical findings underscore the commendable performance of the proposed model, as demonstrated in assessments of the Hotel Arabic-Reviews Dataset, the Book Reviews Arabic Dataset, and the LARB dataset.

Keywords: Transformer; Inductive Transfer; text classification; Arabic dialects; positional encoding; five-polarity

MSC: 68T07

Citation: Baniata, L.H.; Kang, S. Transformer Text Classification Model for Arabic Dialects That Utilizes Inductive Transfer. *Mathematics* **2023**, *11*, 4960. https://doi.org/10.3390/math11244960

Academic Editor: José Antonio Sanz

Received: 9 November 2023
Revised: 8 December 2023
Accepted: 12 December 2023
Published: 14 December 2023

1. Introduction

Text classification, a computational technique, involves discerning and comprehending the emotional sentiment conveyed in a text, be it a sentence, document, or social media message. This process proves invaluable for businesses aiming to glean insights into the perceptions surrounding their brands, products, and services, as gleaned from online interactions with customers. Notably, platforms like Twitter experience a substantial daily surge in user-generated content in Arabic, a trend anticipated to persist with an anticipated rise in user-generated content in the years ahead. Opinions expressed in Arabic are estimated to constitute about five percent of the languages represented on the Internet. Furthermore, Arabic has emerged as one of the most influential languages in the online sphere in recent years. Spoken by a global community of over 500 million individuals, Arabic belongs to the Semitic language family. It serves as the official and administrative language in over 21 countries, spanning from the Arabian Gulf to the Atlantic Ocean. In terms of linguistic structure, Arabic possesses a rich and intricate system compared to English, and its distinctive feature lies in the presence of numerous regional variations. The substantial distinctions between Modern Standard Arabic (MSA) and Arabic dialects (ADs) contribute to the heightened complexity of the language. Additionally, within the realm of Arabic, an intriguing linguistic phenomenon known as diglossia arises, where individuals employ Arabic vernaculars in casual settings and turn to Modern Standard Arabic (MSA)

for formal contexts. As an illustration, residents of Syria adapt their language choice based on the situation, seamlessly switching between MSA and their local Syrian dialects. The Syrian dialect serves as a reflection, mirroring the intricate tapestry of the nation's history, cultural identity, heritage, and shared life experiences. It is worth noting that Arabic dialects exhibit regional diversity, encompassing the Levantine dialects of Palestine, Jordan, Syria, and Lebanon, the Maghrebi dialects prevalent in Morocco, Algeria, Libya, and Tunisia, as well as the Iraqi and Nile Basin dialects spoken in Egypt and Sudan, alongside the Arabian Gulf dialects used in the UAE, Saudi Arabia, Qatar, Kuwait, Yemen, Bahrain, and Oman. Unraveling the sentiments expressed in these various Arabic variations presents a distinctive challenge due to the complex interplay of morphological features, diverse spellings, and the overall linguistic intricacy. Each Arabic-speaking nation takes pride in its unique vernacular, further contributing to the inherent nuances of the language. This linguistic diversity is exemplified by the fact that Arabic texts encountered on social platforms are composed in both Modern Standard Arabic (MSA) and dialectal Arabic, leading to distinct interpretations of the same word. Furthermore, within Arabic dialects (ADs), a notable syntactic concern lies in the arrangement of words. To delve into this matter, it is imperative to pinpoint the verb, subject, and object within an AD sentence. As previously elucidated in the literature review, languages fall into distinct categories based on their sentence structures, such as subject–object–verb (as seen in Korean), subject–verb–object (as in English), verb–object–subject (as in Arabic), and others that allow for flexible phrase order, as is the case in ADs [1]. In AD expressions, this flexible word order imparts crucial information about subjects, objects, and various other elements. Consequently, employing a single-task learning approach and relying solely on manually crafted features proves inadequate for conducting sentiment analysis in Arabic dialects [2]. Furthermore, these disparities in ADs present a formidable challenge for conventional deep learning algorithms and models based on word embeddings. This is because as phrases in ADs grow lengthier, they amass a greater volume of information pertaining to objects, verbs, and subjects, as well as intricate and potentially confounding contextual cues. A drawback of traditional deep learning methods is the loss of input sequence data, leading to a reduction in the performance of the sentiment analysis (SA) model as the length of the input sequence escalates. Furthermore, depending on the context, the root and characters of Arabic words can be in many different forms. In addition, the lack of systematic orthographies is one of the main problems of ADs. This deficiency includes morphological differences between these dialects, which are visible in the use of affixes and suffixes that are not found in MSA.

Furthermore, a multitude of Arabic terms display varying nuances in meaning when employing the same syntax with diacritics. Additionally, the training of sentiment analysis (SA) models rooted in deep learning calls for an extensive corpus of training data, a particularly daunting challenge in the realm of Analytical Dependencies (ADs). ADs, characterized by their lack of structured linguistic elements and resources, pose a significant hurdle for information extraction [3]. As the pool of training data dwindles for ADs, the precision of classification follows suit. Additionally, the majority of tools designed for Modern Standard Arabic (MSA) do not account for the distinctive features of Arabic dialects [4]. Relying on lexical resources, like lexicons, is also considered a less effective approach for Arabic SA due to the vast array of words stemming from various dialects, rendering it improbable to encompass all of them in a lexicon [5]. Moreover, crafting tools and resources tailored specifically for Arabic dialects proves to be a painstaking and time-intensive endeavor [6].

In recent times, there has been a notable upswing in research efforts devoted to sentiment analysis in Arabic vernaculars. These endeavors primarily center around the categorization of reviews and tweets into binary and ternary polarities. The majority of these methodologies [7–12] rely on lexicons, manually crafted attributes, and tweet-specific features, which serve as inputs for machine learning (ML) algorithms. Conversely, alternative approaches adopt a rule-based approach, exemplified by the process of lexicalization. This entails formulating and prioritizing a set of heuristic rules used to classify tweets as either

negative or positive [13]. Moreover, the realm of Arabic sentiment ontology encompasses sentiments with varying degrees of intensity to discern user sentiments and facilitate tweet classification. Additionally, advanced deep learning techniques for sentiment classification [14–18], including recursive auto-encoders, have garnered considerable attention due to their impressive adaptability and robustness in automated feature extraction. Notably, the recently introduced Transformer model [19] has surpassed deep learning models [20] across a range of natural language processing (NLP) tasks, thus piquing the interest of researchers delving into deep learning. The application of the Transformer model, which incorporates a self-attention (SE-A) mechanism to analyze the interactions among words in a sentence, has greatly enhanced the effectiveness of various endeavors in Natural Language Processing (NLP). However, strategies to tackle the challenges of Analytical Dependency Sentiment Analysis (ADs SA) are currently under examination and exploration. Up to this point, no prior research has focused on constructing an ADs text classification model based on self-attention (SE-A) within an Inductive Transfer (INT) framework. Inductive Transfer (INT) enhances comprehension abilities, the quality of the encoder, and the proficiency of text classification compared to a traditional single-task classifier by concurrently addressing related tasks through a shared representation of textual sequences [21]. The primary advantage of Inductive Transfer (INT) lies in its sophisticated approach to utilizing diverse resources for similar tasks. However, the majority of methodologies used in ADs text classification studies lean towards binary and ternary classifications. Hence, our emphasis in this study pertains to the five-polarity AD SA challenge. To the best of our knowledge, no prior investigations have explored the integration of a self-attention (SE-A) approach within an Inductive Transfer (INT) framework for ADs text. Previous techniques addressing this classification dilemma were rooted in single-task learning (STL). In summary, our contributions can be outlined as follows:

- In this article, we present a cutting-edge Transformer model that incorporates Inductive Transfer (INT) for the purpose of conducting analytical dependencies text classification. We have meticulously designed a specialized INT model that makes use of the self-attention mechanism in aiming to exploit the connections between three to five distinct polarities. To achieve this, we employ a transformer encoder, which integrates both self-attention and Feed-Forward Layers (FFLs), thus serving as a foundational layer shared across the tasks. We elucidate the process of simultaneously and interchangeably training on two tasks—ternary and five-class classifications—within the Inductive Transfer (INT) framework. This strategic approach aims to enhance the representation of ADs text for each task, resulting in an expanded spectrum of captured features.
- In this research endeavor, we investigated how altering the number of encoders and utilizing multiple attention heads (AHs) within the self-attention (SE-A) sub-layer of the encoder influences the performance of the proposed model. The training regimen encompassed a range of dimensions for word embedding and utilized a shared vocabulary that was applied across both tasks.

The remainder of this paper is organized as follows. Section 2 presents the literature review. Section 3 discusses the proposed model in detail. Section 4 presents the results of the experiments. Finally, the conclusions of this study are presented in Section 5.

2. Related Work

Research focused on five-level classification tasks in Arabic text classification has received less attention compared to binary and ternary polarity classification tasks. Moreover, the majority of approaches applied to this task rely on traditional machine learning algorithms. For instance, methods based on corpora and lexicons were evaluated using Bag of Words (BoW) features along with a range of machine learning algorithms, including passive-aggressive (PA), Support Vector machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), perceptron, and stochastic gradient descent (SGD) on the Arabic Book Review dataset [22]. In a related study using the same dataset, [23] explored the effects of stemming

and balancing the BoW features using various machine learning algorithms. They discovered that stemming led to a decrease in performance. Another strategy presented in [24] proposed a divide-and-conquer approach to tackle ordinal-scale classification tasks. Their model was organized as a hierarchical classifier (HC) in which the five labels were divided into sub-problems. They noted that the HC model outperformed a single classifier. Building on this foundation, several hierarchical classifier configurations were suggested [25]. These configurations were compared with machine learning classifiers, such as SVM, KNN, NB, and DT. The experimental outcomes indicated that the hierarchical classifier enhanced performance. Nevertheless, it should be noted that many of these configurations demonstrated reduced performance. Another investigation [26] delved into various machine learning classifiers, encompassing LR, SVM, and PA, using n-gram characteristics within the Book Reviews in Arabic Dataset (BRAD). The findings demonstrated that SVM and LR achieved the most impressive results. Similarly, [27] assessed a range of sentiment classifiers, including AdaBoost, SVM, PA, random forest, and LR, on the Hotel Arabic-Reviews Dataset (HARD). Their findings highlighted that SVM and LR exhibited superior performance when paired with n-gram features. The previously mentioned methodologies conspicuously overlook the incorporation of deep learning techniques for the five-tier polarity classification tasks in Arabic Sentiment Analysis. Furthermore, the majority of approaches targeted at these five polarity tasks rely on traditional machine learning algorithms that hinge on the feature engineering process, a method deemed time consuming and laborious. Additionally, these approaches are founded on single-task learning (STL) and lack the capability to discern the interrelationships between various tasks (cross-task transfer) or to model multiple polarities simultaneously, such as both five and three polarities.

Other research endeavors have employed Inductive Transfer (INT) to address the challenge of five-point Sentiment Analysis (SA) classification tasks. For example, in [28], an Inductive Transfer model centered on a recurrent neural network (RNN) was proposed, which simultaneously handled five-point and ternary classification tasks. Their model comprised Bidirectional Long Short-Term Memory (Bi-LSTM) and multilayer perceptron (MLP) layers. Similarly, in [29], the interplay between five-polarity and binary sentiment classification tasks was leveraged by concurrently learning them. The suggested model featured an encoder (LSTM) and decoder (variational auto-encoder) as shared layers for both tasks. The empirical results indicated that the INT model bolstered performance in the five-polarity task. The introduction of adversarial multitasking learning (AINT) was first presented in [21]. This model incorporated two LSTM layers as task-specific components and one LSTM layer shared across tasks. Additionally, a Convolutional Neural Network (CNN) was integrated with the LSTM, and the outputs of both networks were merged with the shared layer output to form the ultimate sentence latent representation. The authors determined that the proposed INT model elevated the performance of five-polarity classification tasks and enhanced encoder quality. Although the INT strategies delineated above have been applied in English, a conspicuous dearth exists in the application of Inductive Transfer and deep learning techniques to five-polarity Arabic Sentiment Analysis. Existing studies for this task predominantly rely on single-task learning with machine learning algorithms. Therefore, there remains ample room for improvement in the effectiveness of current Arabic Sentiment Analysis approaches for the five polarities, which still stand at a relatively modest level. Further investigations have leveraged deep learning techniques in Sentiment Analysis (SA) across a diverse array of domains encompassing finance [30,31], evaluations of movies [32–34], Twitter posts related to weather conditions [35], feedback from travel advisors [36], and recommendation systems for cloud services [37]. As highlighted by the authors in [35], text attributes were automatically extracted from diverse data sources. Weather-related knowledge and user information were transformed into word embedding through the application of the word2vec tool. This approach has been adopted in multiple research studies [30,38]. Numerous papers have employed sentiment analysis based on polarity using deep learning techniques applied to Twitter posts [39–42]. The researchers elucidated how the employment of deep learning methodologies led to

an augmentation in the precision of their distinct sentiment analyses. While the majority of these deep learning models have been implemented for English text, there are a few models that handle tweets in other languages, such as Persian [38], Thai [41], and Spanish [43]. The researchers conducted analyses on tweets utilizing various models tailored for polarity-based Sentiment Analysis, including Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), hybrid models [42], and Support Vector Machines (SVMs) [44].

Various methods have been suggested to identify fake information, with most of them concentrating solely on linguistic attributes and disregarding the potential of dual emotional characteristics. Luvembe et al. [45] present an efficient strategy for discerning false information by harnessing dual emotional attributes. To accomplish this, the authors employ a stacked bi-GRU layer to extract emotional traits and represent them in a feature vector space. Furthermore, the researchers implement a profound attention mechanism atop the Bi-GRU to enhance the model's ability to grasp vital dual-emotion details. Lei et al. [46] present MsEmoTTS, a multi-scale framework for synthesizing emotional speech that aims to model emotions at various tiers. The approach follows a standard attention-based sequence-to-sequence model and integrates three crucial components: the global-level emotion-representation module (GM), the utterance-level emotion-representation module (UM), and the local-level emotion-representation module (LM). The findings demonstrate that MsEmoTTS surpasses existing methods relying on reference audio and text-based approaches for emotional speech synthesis, excelling in both emotions transfer and emotion prediction tasks. Li et al. [47] introduce a fresh methodology termed "Attributed Influence Maximization based on Crowd Emotion" (AIME). The main goal is to scrutinize the influence of multi-dimensional characteristics on the dissemination of information by incorporating user emotion and group attributes. The findings demonstrate that AIME surpasses heuristic approaches and yields results comparable to greedy methods while significantly enhancing time efficiency. The experimental results affirm its superiority over traditional methods, highlighting its potential as an efficient and potent tool for comprehending the dynamics of information dissemination. A plethora of scholarly articles have increasingly relied on Twitter datasets as a fundamental source for constructing and refining models for sentiment analysis. For example, Vyas et al. [48] introduced an innovative hybrid framework that blends a lexicon-based approach with deep learning techniques to scrutinize and categorize the sentiments conveyed in tweets pertaining to COVID-19. The central aim was to automatically discern the emotions conveyed in tweets about this subject matter. To accomplish this, the authors employed the VADER lexicon method to identify positive, negative, and neutral sentiments, which were then used to tag COVID-19-related tweets accordingly. In the categorization process, a range of machine learning (ML) and deep learning (DL) techniques were utilized. Furthermore, Qureshi et al. [49] introduce a model designed to assess the polarity of reviews written in Roman Urdu text. The research employed nine diverse machine learning algorithms, including Naïve Bayes, SVM, Logistic Regression, K-Nearest Neighbors, ANN, CNN, RNN, ID3, and Gradient Boost Tree. Among these algorithms, Logistic Regression showcased the highest efficacy, demonstrating superior performance with testing and cross-validation accuracies of 92.25% and 91.47%, respectively. Furthermore, Alali [2] introduced a multitasking approach termed the Multitask learning Hierarchical Attention Network, which was engineered to enhance sentence representation and enhance generalization. The outcomes of the experiments underscore the outstanding effectiveness of this proposed model.

By utilizing the BRAD, HARD, and LARB datasets with five-point scales for investigating ADs, we evaluated the T-TC-INT model specifically tailored for this purpose against the most current standard techniques. Initially, Logistic Regression (LR) was introduced in [26] by employing unigrams, bi-grams, and TF-IDF, and it was subsequently applied to the BRAD dataset. Similarly, Logistic Regression (LR) was initially recommended in [27] using unigrams, bi-grams, and TF-IDF, and it was then put into action on the HARD dataset. Our proposed T-TC-INT model has also undergone a comparative assessment utilizing the

LABR datasets. These established methods encompass the following: first, the Support Vector Machine (SVM), which employs a Support Vector Machine classifier with n-gram characteristics, as advocated in [23]. Second, the Multinomial Naïve Bayes (MNB), which implements a multinomial Naïve Bayes approach with bag-of-words attributes, as outlined in [22]. Third, the Hierarchical Classifiers (HCs), a model utilizing hierarchical classifiers, is constructed based on the divide-and-conquer technique introduced by [24]. Finally, Enhanced Hierarchical Classifiers (HC(KNN)), which is a refined iteration of the hierarchical classifiers model still rooted in the divide-and-conquer strategy, as elucidated by [25]. In recent developments, tasks in Natural Language Processing (NLP) have achieved remarkable proficiency by harnessing the power of the bi-directional encoder representation from transformers, known as BERT [50]. The AraBERT [51], an Arabic pre-trained BERT model, underwent training on three distinct corpora: OSIAN [52], Arabic Wikipedia, and the MSA corpus, encompassing a staggering 1.5 billion words. We conducted a comparative evaluation between the proposed ST-SA system for ADs and AraBERT [51], which boasts 786 latent dimensions, 12 attention facets, and a composition of 12 encoder layers.

3. The Proposed Transformer-Based Text Classification Model for Arabic Dialects That Utilizes Inductive Transfer

We designed a text classification model employing self-attention (SE-A) and Inductive Transfer (INT) for classifying text in ADs. The goal of utilizing INT (Inductive Transfer) is to improve the performance of Arabic sentiment analysis, which involves classifying ADs into five-point scales, by leveraging the relationship between the AD sentiment analysis tasks (in both five and ternary polarities). Our approach, the Transformer text classification T-TC-INT model that utilizes Inductive Transfer (INT), is based on the Transformer model recently introduced by Vaswani et al. [19]. Inductive Transfer (INT) proves more advantageous than single-task learning (STL), as it employs a shared representation of various loss functions while simultaneously addressing sentiment analysis tasks (with three and five polarities) to enhance the representation of semantic and syntactic features in ADs text. Knowledge gained from one task can benefit others in learning more effectively. Additionally, a significant advantage of INT is its ability to tap into resources developed for similar tasks, thereby enhancing the learning process and increasing the amount of useful information available. Sharing layers across associated tasks enhances the model's ability to generalize, learn quickly, and comprehend what it has learned. Furthermore, by leveraging domain expertise observed in the training signals of interconnected tasks as an inductive bias, the INT method facilitates prompt transfers that improve generalization. This Inductive Transfer aids in increasing the accuracy of generalization, the speed of learning, and the clarity of the learned models. A learner engaged in multiple interconnected tasks simultaneously can use these tasks as an inductive bias for one another, resulting in a better grasp of the domain's regularities. This approach allows for a more effective understanding of sentiment analysis tasks for ADs, even with a small amount of training data. Furthermore, Inductive Transfer can concurrently capture the significant interrelationships between the tasks being learned. As illustrated in Figure 1, our proposed T-TC-INT approach features a unique architecture based on the SE-A, INT, shared vocabulary, and 1D_GlobalAveragepooling layer. The T-TC-INT model we propose is designed to handle diverse classification tasks, encompassing both ternary and five-polarity classification tasks, and it learns them simultaneously. By incorporating a shared Transformer block (encoding layer), we enable the transfer of knowledge from the ternary task to the five-point task as part of the learning process. This leads to an enhancement in the learning capabilities of the current task (the five-point task). The encoder within our proposed T-TC-INT model, as depicted in Figure 1, consists of a range of distinct layers. Each layer is further divided into two sub-layers: self-attention (SE-A) and position-wise feed-forward (FNN).

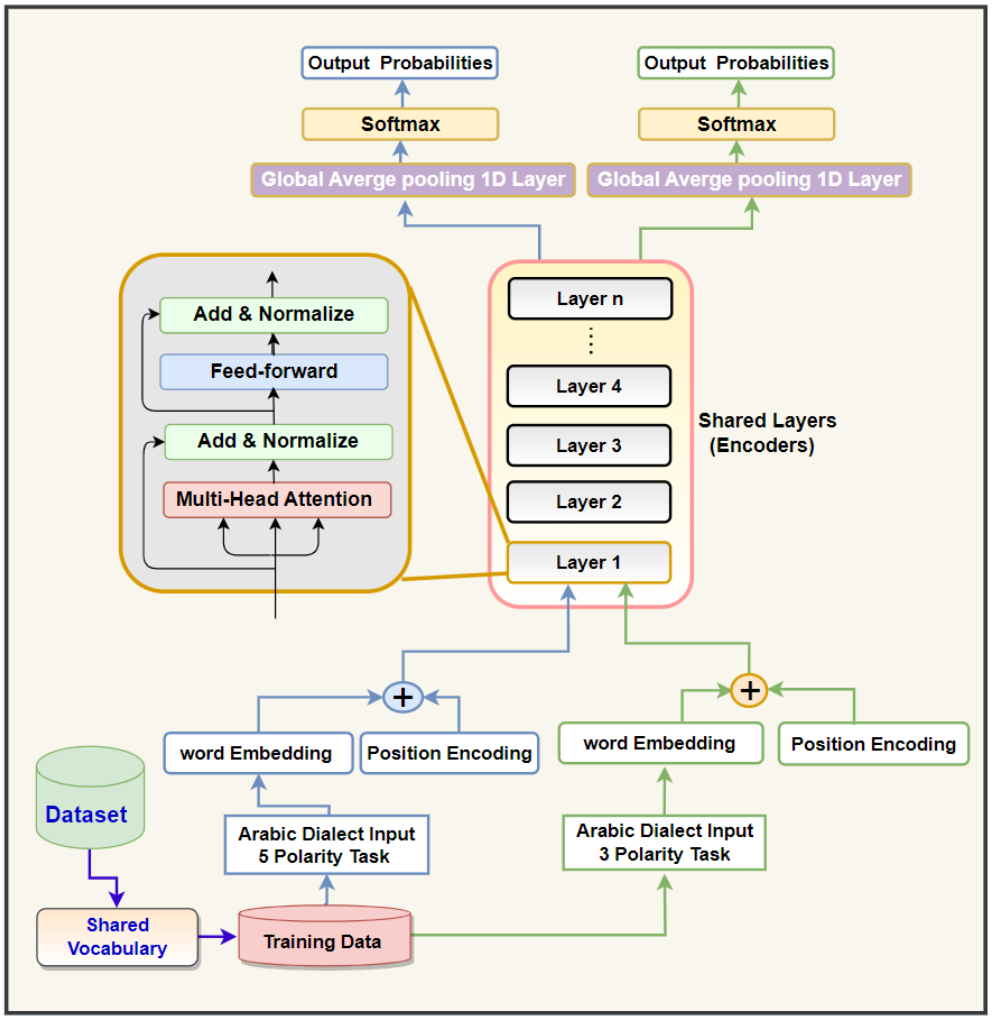

Figure 1. The architecture of the suggested T-TC-INT model for Arabic Dialects.

In this suggested structure, the encoder employs the SE-A and FNN sub-layers to classify ADs of varying lengths without resorting to RNN or CNN algorithms. The model we propose utilizes these sub-layers to encode the input phrases from ADs, thus generating an intermediate representation. The encoder in the suggested model corresponds the input sentence $(x_1, \ldots\ldots, x_n)$ into a vector representation sequence $Z = (z_1, \ldots\ldots, z_n)$. By considering Z, the T-TC-INT model generates the sentiment of an AD sentence. Word embeddings, a form of word representation in the realm of natural language processing (NLP), are designed to encapsulate the semantic associations among words. They manifest as numerical vectors that depict words within a continuous, compact space, where words sharing akin meanings or contextual relevance are positioned in closer proximity. Word embeddings have evolved into a cornerstone of numerous NLP applications, including machine translation and sentiment analysis. In contrast, traditional approaches for representing words in NLP, such as one-hot encoding, represent words as sparse binary vectors. In this scheme, each word receives a distinct index and is essentially treated as a standalone entity. This representation falls short of capturing the underlying semantic connections between words, thereby posing challenges for machine learning models in grasping word meanings in context. Word embeddings, conversely, capture these semantic links by adhering to the distributional hypothesis, which posits that words found in similar contexts tend to share similar meanings. By mapping words to dense vectors within a continuous space, word embeddings possess the ability to capture subtleties in meaning, synonymy, analogy, and more. This empowers machine learning models to acquire a deeper comprehension of word semantics and their interrelations. What sets word embeddings apart is their contextual

richness. The embedding of a word can adapt to varying contexts, and it changes based on the surrounding words. This dynamic feature permits the representation to flexibly adjust and align with different contextual scenarios. The embedding layer in the encoder converts the input tokens (source tokens in the encoder) into a vector of dimension d_{model}. When the information needed for token proximity is interpreted in the SE-A sub-layer, the information for the token's position is embedded through positional encoding (PE). In particular, the PE is a matrix that represents the position details of the tokens in the input sentence, and the suggested T-TC-INT system integrates the PE into the embedding matrix of the input tokens. Each part of the PE is determined utilizing *sine* and *cosine* function equations with unstable frequencies.

$$PE\ (pos, 2i) = \sin\left(pos/10,000^{2i/d_{model}}\right), \tag{1}$$

$$PE\ (pos, 2i+1) = \cos\left(pos/10,000^{2i/d_{model}}\right), \tag{2}$$

where the *pos* indicates the position of each input token, i represents the dimension of each element, and d_{model} represents the embedding dimension of the input tokens. The embedding matrix combined with the PE is fed to the encoder's first layer. The encoder subnetwork consists of L similar layers, where L is set to different numbers (12, 8, 2, and 4). Each encoding layer consists of two layers: an SE-A sub-layer and an FNN sub-layer. A residual connection method and a layer normalization unit (Layer Norm) are utilized on every sub-layer to support training and improve performance. The output of every layer l (H_e^l) is calculated as follows:

$$S_e^l = LayerNorm\left(MHA\left(H_e^{l-1},\ H_e^{l-1}, H_e^{l-1}\right) + H_e^{l-1}\right), \tag{3}$$

$$H_e^l = LayerNorm\left(FFN\left(S_e^l\right) + S_e^l\right), \tag{4}$$

where S_e^l is the output from the SE-A sub-layer calculated based on the input sentence representation of the prior encoding layer $(l-1)$.

Self-Attention (SE-A)

The self-attention mechanism (SE-A) is a key component of the seq-2seq design, and it is utilized to solve an array of sequence-generation problems, such as NMT and document summarization [53]. Figure 2 shows how the T-TC-INT model for ADs accomplishes the scale-dot product attention function. This function uses three vectors as inputs: queries Q, values V, and keys K. It describes the given query and key–value pairs as a weighted sum of values. The weights reveal the relationship between each query and the key. The following is an explanation of the attention method.

$$Attention(Q, K, V) = softmax(\alpha)V \tag{5}$$

$$\alpha = score(Q, K) \tag{6}$$

$$score(Q, K) = \frac{Q \times K^T}{\sqrt{d_k}} \tag{7}$$

where $k \in R^{J \times d_k}$ is the key, $V \in R^{J \times d_v}$ is the value, and $Q \in R^{Z \times d_k}$ is a query. Z and J are the lengths of the sequences expressed as Q and K, respectively. d_k and d_v are the dimensions of the value and key vectors, respectively. The query dimension is expressed as d_k to perform the dot product calculation. We divided $Q \times K^T$ by $\sqrt{d_k}$ to measure the output of the product operation, thereby maintaining the calculation of Vaswani et al. [19]. The overall attention weight distribution was obtained by applying the $softmax(.)$ operation

to the attention score $\alpha \in R^{Z \times J}$. For better performance, the transformer architecture uses SE-A, which comprises N_h (number of head attentions) measured dot product attention operations. Provided with the Q, K, and V, the SE-A computation is as follows:

$$MHA(Q, K, V) = O, \tag{8}$$

$$O = HW_o, \tag{9}$$

$$H = concat\left(head_1, head_2, \ldots\ldots, head_{N_h}\right) \tag{10}$$

$$head_h = Attention\left(QW_h^Q, KW_h^k, VW_h^v\right) \tag{11}$$

where QW_h^Q, KW_h^k, and VW_h^v are the projections of the query, key, and value vectors, respectively, for the h^{th} head. These projections were made using the metrics $W_h^Q \in R^{d_{model} \times d_k}$, $W_h^k \in R^{d_{model} \times d_k}$, and $W_h^v \in R^{d_{model} \times d_v}$. The inputs to the MHA(.) are $K \in R^{J \times d_{model}}$, $V \in R^{J \times d_{model}}$, and $Q \in R^{Z \times d_{model}}$. $head_h \in R^{J \times d_v}$ is the output of the measured dot product calculation for the h^{th} head. The N_h measured dot product operation is combined using the concatenation function $concat(.)$ to generate $H \in R^{Z \times (N_h . d_v)}$. Eventually, the output $O \in R^{Z \times d_{model}}$ is obtained from the projections of H using the weight matrix $W_o \in R^{(N_h . d_v) \times d_{model}}$. The SE-A contains the same number of parameters as the vanilla attention.

$$\text{if } d_k = d_v = \frac{d_{model}}{N_h} \tag{12}$$

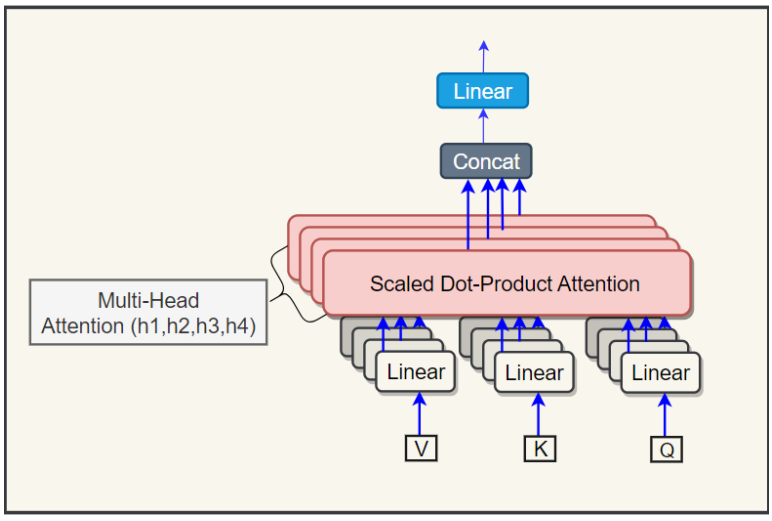

Figure 2. The structure of the self-attention (SE-A) sub-layer.

The T-TC-INT model leverages the self-attention (SE-A) mechanism, thus enabling it to discern the relative importance of each word within a sentence and capture subtle sentiments and their varying degrees of intensity. Furthermore, the model incorporates a self-attention mechanism (SE-A) that allows it to contextualize words in relation to their neighboring words, thus leading to more refined sentiment predictions. The T-TC-INT model has undergone meticulous training on a diverse dataset to ensure its efficacy across a spectrum of Arabic text styles and domains. Consequently, the proposed approach for text classification in ADs using Transformer (T) and Inductive Transfer (INT) provides a more precise representation of sentiments, which effectively addresses the challenge of varying intensities within the Arabic text classification spectrum.

4. Practical Experiments

Several practical experiments were performed in order to assess the T-TC-INT model for Arabic dialect ADs. The ADs classification performance of the proposed T-TC-INT model was assessed.

4.1. Data

The proposed model was trained using three prominent benchmark Arabic dialect datasets. The initial dataset, HARD [27], consisted of reviews sourced from various reservation websites, which were subsequently categorized into five distinct groups. The second dataset, BRAD [26], and the third dataset, LARB [22], were also utilized. All three datasets (BRAD, HARD, and LARB) employed in this research were analyzed at the review level. BRAD reviews were collected from the Goodreads website and organized into five scales. Detailed class distributions for the HARD, BRAD, and LARB datasets are presented in Tables 1–3, respectively. A pre-processing phase was performed for all of the datasets used in this research project. For instance, all sentences underwent pre-processing. This involved employing sentence breakers to segment reviews into individual sentences. Additionally, any English letters, non-Arabic words and characters, diacritics, hashtags, punctuation, and URLs were completely removed from all ADs texts. Orthographic normalization was applied to all ADs texts. For instance, (أ) letters were normalized to (ا), (ي) letters were normalized to (ى), and (ؤ) letters were normalized to ((و. Furthermore, emoticons were replaced with their corresponding explanations, and elongated words were preprocessed. To prevent potential overfitting, an early stopping mechanism with a patience parameter set to three epochs was employed. Additionally, when evaluating the proposed T-TC-INT model for ADs, the model checkpoint mechanism was utilized to save the most optimal weights of the suggested model. In addition to their partitioning for training and testing, the HARD, BRAD, and LARB datasets offer valuable insights into the distribution of polarities across their samples. The HARD dataset, comprising a total of 409,562 samples, was divided into five polarities, each representing distinct sentiments or attitudes. With 80% of the dataset allocated for training (327,649 samples) and 20% for testing (81,912 samples), these proportions provided a comprehensive representation of the different polarities within both subsets. Similarly, the BRAD dataset, encompassing 510,598 samples, was partitioned into 80% (408,478 samples) for training and 20% (101,019 samples) for testing. Also, the LARB dataset, encompassing 63,257 samples, was partitioned into 80% (50,606 samples) for training and 20% (12,651 samples) for testing. This partitioning scheme ensured that the five polarities were adequately represented in both the training and testing phases, thus enabling the models to grasp the nuances of sentiment variation and effectively generalize their understanding to unseen data. Biases can have a major influence on sentiment analysis model performance. When biases exist in the data that are used to train these model, they can lead to distorted results. To address the issue of biases and how to select data for the proposed T-TC-INT text classification model for ADs, we considered four steps:

- We ensured that the training data included a variety of sources and covered a wide range of demographics, geographical areas, and social circumstances; this allowed for decreasing biases and provided a more comprehensive and balanced dataset.
- We ensured that the sentiment labels in the training data were balanced across all demographics groups and opinions; this assisted in minimizing overgeneralization and biases caused by an imbalance in sentiment examples.
- We established clear labeling guidelines that explicitly instructed human annotators to be impartial and avoid injecting their own biases into the sentiment labels; this can help maintain consistency and minimize biases.
- We performed a detailed analysis of the training data to identify possible biases. This can include checking demographic imbalances, serotype reinforcement, or underrepresented groups. Once detected, we took suitable steps to address these bi-

ases, such as data augmentation, oversampling of underrepresented groups, and pre-processing techniques.

Table 1. Statistics for HARD imbalanced 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 2. Statistics for BRAD imbalanced 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 3. Statistics for LARB 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. Settings of the Proposed T-TC-INT Model

The proposed T-TC-INT text classification model was created using Tensor-Flow [54], Keras [55], and scikit-learn [56] frameworks. The experiments for all Arabic dialects (ADs) classification tasks, encompassing both three and five polarities, were conducted employing diverse parameter configurations. These included six values for the word-embedding dimension: 512, 300, 256, 128, 64, and 50. Additionally, the attention heads were varied with options of 12, 10, 8, 6, 4, and 2. The position-wise Feed-Forward Neural Network (FNN) featured filters with dimensions of 512, 300, 256, 128, 64, and 50. The recently devised model incorporates a shared encoder sub-network with layers 3, 2, and 1, as well as a one-dimensional average pooling layer tailored for each classification task (ternary and five polarities) on ADs.

4.3. Training Mechanism of the Proposed T-TC-INT Model

Inductive Transfer models employ two primary training methodologies: joint training and alternative training. Joint training entails simultaneously training a model on multiple tasks, thus allowing for the sharing of information and the acquisition of representations that benefit all tasks. This approach capitalizes on the interdependencies between tasks to enhance overall performance. Conversely, alternative training focuses on training the model on tasks one at a time while cycling through them iteratively. This enables the model to give concentrated attention to each task, potentially leading to improved performance for each task in isolation. Both approaches have their merits and drawbacks. Joint training can lead to superior generalization across tasks, while alternative training may excel in tasks with significant disparities in data distribution or complexity. The selection between these strategies hinges on the specific characteristics of the tasks at hand and the desired trade-offs in performance and efficiency. Ultimately, the choice of training methodology plays a pivotal role in shaping the effectiveness and adaptability of Inductive Transfer models. The suggested system adeptly manages both three-category and five-category classification tasks. For instance, during the training process on the HARD dataset, the proposed T-TC-INT text classification system seamlessly transitions between instructing the model on the five-category and three-category classification tasks. We examined two different strategies for training the model: alternating [23,57] and joint learning. In

our Inductive Transfer configuration, we applied the loss function and optimizer for each task sequentially. This implies that the training procedure begins with the three-category classification task for a specified number of epochs before shifting to the five-category classification task. The primary goal in training both tasks was to minimize the categorical cross-entropy. The proposed T-TC-INT text classification model underwent training for 12 epochs, implementing an early stopping mechanism set to activate after 2 epochs, with a batch size of 90. We followed standard protocols for the HARD, BRAD, and LARB datasets, dividing them into an 80% training subset and a 20% testing subset. We opted for the Nadam optimizer to guide each task within the proposed T-TC-INT text classification model. We employed a sentence breaker to segment reviews into sentences, with maximum sentence lengths set at 80 for HARD, 50 for BRAD, and 80 for LARB. The class weights methodology was not integrated into the proposed model [24]. Prior to each epoch, the training data underwent a random shuffling process. Further details on hyper-parameters can be located in Section 4.4.

4.4. Results

Numerous practical trials were conducted using the recommended T-TC-INT text classification (TC) system tailored for Arabic dialects. The T-TC-INT text classification (TC) system was put through training with varying configurations of attention heads (AHs) in the self-attention (SE-A) sub-layer, as well as different numbers of encoders, to determine the most effective system structure. Diverse word-embedding dimensions were employed in the training of the proposed T-TC-INT text classification system. This study delved into the repercussions of training the proposed system using two multitasking methodologies—jointly and alternately—to gauge its performance. The efficacy of sentiment analysis in the proposed system was assessed using an automated accuracy metric. This section presents the evaluation of the proposed T-TC-INT text classification system for the task of classifying sentiments into five categories for ADs. The outcomes of the practical experiments on the datasets HARD, BRAD, and LARB are detailed in Tables 4–6, respectively. As elucidated in Tables 4 and 7, the T-TC-INT text classification system attained an accuracy of 81.83%, precision of 81.27%, recall of 80.04%, and an F-Score of 80.91% on the imbalanced HARD dataset. In this setup, the number of attention heads was two, the encoding layers numbered two, and the word-embedding dimension stood at 300. This commendable accuracy was achieved due to the positive impact of employing the Inductive Transfer (INT) framework and the self-attention (SE-A) approach to right-to-left texts, such as ADs. In comparison to the top-performing LR [58] system on the HARD dataset, the results obtained by the proposed T-TC-INT text classification model surpassed it, exhibiting an accuracy difference of 5.79%. Furthermore, the proposed model outperforms AraBERT [51], with a difference in accuracy of 0.98%. Hence, simultaneous learning tasks bolstered the volume of usable data, and the risk of overfitting was mitigated [59]. The suggested system demonstrated proficiency in capturing both syntactic and semantic features, effectively discerning sentiments within ADs sentences.

Table 4. Results for the T-TC-INT model on the HARD dataset for the five-polarities classification task, where W-E-D is the word embedding dimension, FS is the filter size, EL is the encoding layer, and AH is the number of attention heads.

W-E-D	FS	EL	AH	Accuracy (5-Polarity)	Precision	Recall	F-Score
256	256	2	8	74.53%	73.65%	71.50%	73.81%
256	256	2	12	81.66%	80.47%	79.90%	78.90%
512	512	1	8	80.24%	80.01%	79.85%	79.82%
256	256	1	8	81.52%	81.20%	79.84%	80.02%
256	256	2	10	80.14%	79.98%	78.44%	79.01%
300	300	2	2	81.83%	81.27%	80.04%	80.91%

Table 5. Results for the proposed T-TC-INT model on the BRAD dataset for the five-polarities classification task.

W-E-D	FS	EL	AH	Accuracy (5-Polarity)	Precision	Recall	F-Score
64	164	3	2	61.07%	60.85%	59.43%	59.17%
64	164	2	4	61.01%	60.78%	60.63%	59.67%
50	128	1	2	56.92%	55.39%	54.94%	53.84%
128	128	2	6	61.73%	60.86%	60.77%	61.40%
128	128	2	8	61.32%	61.09%	60.00%	61.10%
200	200	2	6	51.48%	49.86%	50.14%	50.27%

Table 6. Results for the T-TC-INT model on the LARB dataset for the five-polarities classification task.

W-E-D	FS	EL	AH	Accuracy (5-Polarity)	Precision	Recall	F-Score
64	164	3	2	63.27%	62.33%	62.94%	62.87%
256	164	4	4	72.92%	71.30%	72.16%	71.00%
256	128	2	2	70.37%	69.98%	70.03%	69.89%
200	128	2	2	78.13%	77.80%	76.98%	77.35%
128	128	2	6	67.57%	66.57%	65.84%	66.15%
200	200	2	6	75.68%	74.70%	74.87%	74.29%

Table 7. The performance of the proposed T-TC-INT model compared with benchmark approaches on the HARD imbalanced dataset.

Model	Polarity	Accuracy	F-Score
LR [58]	5	76.1%	75.9%
AraBERT [51]	5	80.85%	77.88%
Proposed T-TC-INT Model	**5**	81.83%	80.91%

Furthermore, the proposed T-TC-INT text classification system demonstrated its superior effectiveness on the imbalanced BRAD dataset. As outlined in Table 5, the suggested model achieved an accuracy of 61.73%, precision of 60.86%, recall of 60.77% and an F-score of 61.40%. This was observed when employing six attention heads (AHs), two encoding layers, and a word-embedding dimension of 128. As elucidated in Table 8, the recommended T-TC-INT text categorization system surpassed the Logistic Regression (LR) method put forward by [26] with a notable accuracy margin of 14.03%. It also outperformed the AraBERT model [51] with an accuracy difference of 0.88%. Additionally, the integration of a self-attention-based shared encoder with two distinct global-average pooling one-dimensional layers (one for each classification task) empowers the proposed model to capture a comprehensive representation, encompassing the preceding, subsequent, and local contexts of any position within a sentence. Moreover, the introduced T-TC-INT model, detailed in Table 6, exhibited outstanding performance on the demanding LARB imbalanced dataset. This innovative model attained a commendable accuracy of 78.13%, precision of 77.80%, recall of 76.98% and an F-score of 77.34%, surpassing alternative AP methodologies. Specifically, employing a defined setup featuring two attention heads (AHs), two encoding layers, a filter size of 128, and a word-embedding dimension of 200, the proposed system demonstrated its efficacy. This accomplishment highlights the resilience of the INT-MHA SA model in tackling the intricacies of sentiment analysis within the imbalanced dataset scenario. As shown in Table 9, the INT-MHA SA system outperformed various alternative methods, showcasing its superiority by significant margins. For example, it displayed an impressive accuracy difference of 27.83% when compared to the SVM [23] model, a substantial 33.13% accuracy gap over the MNP [22] model, a remarkable 20.33% accuracy lead over

the HC(KNN) [24] model, as well as a noteworthy 19.17% accuracy difference compared to AraBERT [51]. Additionally, the proposed model even outperformed HC(KNN) [25] by an accuracy margin of 5.49%.

Table 8. The performance of the proposed T-TC-INT model compared with benchmark approaches on the BRAD imbalanced dataset.

Model	Polarity	Accuracy	F-Score
LR [26]	5	47.7%	48.9%
AraBERT [51]	5	60.85%	58.79%
Proposed T-TC-INT Model	5	61.73.%	61.40%

Table 9. The performance of the proposed T-TC-INT model compared with benchmark approaches on the LARB 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 [51]	5	58.96%	55.88%
HC(KNN) [25]	5	72.64%	74.82%
Proposed T-TC-INT Model	5	78.13%	77.80%

Joint training in deep learning is a methodology where a single neural network model is trained to tackle multiple interrelated tasks simultaneously. Instead of training distinct models for each task, joint training enables the model to exchange and acquire representations that can be advantageous for all tasks. This can result in enhanced adaptability, heightened efficiency, and potentially even elevated performance for each specific task. Imbalanced data denote a scenario in which the distribution of categories within a dataset is uneven. In certain instances, one or more categories may possess noticeably fewer instances compared to others. Imbalanced data can pose difficulties for deep learning models as they may become skewed towards the predominant category, resulting in subpar performance for the minority categories. The assessment results highlight that the suggested T-TC-INT system, employing both joint and alternate learning, exhibited exceptional efficacy. Alternate training surpassed joint learning, attaining accuracies of 79.07% and 77.59% in the imbalanced HARD dataset, and 61.35% and 60.98% in BRAD, respectively, as outlined in Table 10. A comparison with the benchmark techniques unveiled that alternate training in the five-point classification setting can facilitate the extraction of more comprehensive feature representations within the text sequence compared to the single-task learning approach. These findings underscore that alternate learning proves more apt for addressing intricate SA tasks and has the capacity to comprehend and generate a more robust latent representation in nuanced tasks for ADs SA.

Table 10. Performance of joint and alternate training for five-polarity classification.

T-TC-INT-Training Method	HARD (Imbalance) Accuracy	BRAD (Imbalance) Accuracy
Alternately	79.07%	61.35%
Jointly	77.59%	60.98%

The distinction in effectiveness between the two methodologies lies in how alternate learning is influenced by the volume of data in each task. Tasks with larger datasets lead to richer shared layers. Conversely, joint learning may exhibit bias if one task's dataset

significantly outweighs the other. Hence, for tasks in ADs text classification (TC), alternative training methods are recommended. Consider a scenario where we have two distinct datasets for separate tasks, such as machine translation tasks involving translation from AD to MSA and from MSA to English [1]. The efficacy of each task can be heightened by constructing a network in an alternate configuration without the need for additional training data [60]. Additionally, other tasks of relevance can bolster the efficiency of five-point classification. The significance of the enhancements in our proposed model's performance can be attributed to several factors. Surpassing a state-of-the-art model like AraBERT and LR is a notable feat in itself, given AraBERT's widely acknowledged effectiveness in Arabic language processing tasks. By outperforming AraBERT on the same datasets, our proposed model demonstrates its heightened accuracy and F-score in handling Arabic dialects. Furthermore, even a slight uptick in accuracy holds significance as it contributes to elevating the overall performance of models for processing Arabic dialects. Even minor improvements can have practical implications, such as refining the accuracy of sentiment analysis, information retrieval, or other NLP applications tailored for Arabic dialects. When contrasting the proposed Transformer Text classification model that utilizes an Inductive Transfer (T-TC-INT) model with AraBERT [51], discernible differences emerge concerning the programming exertion and developmental expenses. While the T-TC-INT model is groundbreaking and proficient in attaining outstanding text classification outcomes, it may entail a marginally elevated programming effort due to its incorporation of multi-task learning and a multi-head attention architecture. This intricacy might result in a modest augmentation of the time and exertion required for development in comparison to the more straightforward AraBERT model. Nevertheless, it is crucial to highlight that the heightened accuracy and adaptability of the T-TC-INT model have the potential to rationalize the supplementary effort invested in its creation. Conversely, AraBERT, functioning as a pre-trained BERT-based model, might present a comparatively more straightforward implementation process, thereby potentially diminishing development time and associated costs. Ultimately, the choice between these two models hinges on the specific requisites of the project, the resources available, and the balance between accuracy and developmental effort.

4.5. Impact of Attention Head (AH) Number V in Self-Attention Sub-Layer

As depicted in Tables 4–6, the efficacy of the proposed T-TC-INT text classification system, derived from varied input representations sourced from the self-attention layer, underscores the significance of this model in the five-polarity classification task. Here, V denotes the number of attention heads (AHs) within the encoding layer of the recommended T-TC-INT text classification system. The suggested system underwent training utilizing a range of V values: 2, 4, 6, 8, 10, and 12. As evidenced in Tables 4–6, there was a discernible shift in the accuracy metrics for the datasets HARD, BRAD, and LARB.

4.6. Impact of Length of Input Sentence

Enhancing the classification of lengthy sentences hinges on acquiring comprehensive contextual knowledge and dependencies across tokens in input phrases. To facilitate this, sentences of comparable length (in terms of source tokens) were grouped together, following a methodology similar to that of Luong et al. [61]. Given the extensive nature of the HARD dataset, a five-polarity classification task was undertaken to assess the sentiment analysis (SA) performance specifically for lengthy sentences. The evaluation in this section is based on distinct length categories: <10, 10–20, 20–30, 30–40, 40–50, and >50. An automated accuracy metric was computed for the output generated by the T-TC-INT system. As depicted in Table 11, the effectiveness of the recommended T-TC-INT text classification model saw an improvement as the input sentence length increased, especially for tokens comprising 30 to 40 words and those exceeding 50 words, registering accuracy scores of 79.03 and 81.83, respectively. Leveraging multi-task learning, a self-attention mechanism, and harnessing word units as input features for the (SE-A) sub-layer, the proposed system

gains contextually relevant knowledge and dependencies within tokens, regardless of their position within the ADs input phrase. However, shorter sentences containing between 10 and 20 words, as well as those with fewer than 10 words, demonstrated lower efficacy for the proposed model. Additionally, the system's performance was notably lower for sentences with fewer than 10-word tokens, exhibiting an accuracy of 76.88. As illustrated in Tables 12 and 13, the performance of the proposed T-TC-INT text classification model saw an improvement as the input sentence length increased, especially for tokens exceeding 50 words, registering accuracy scores of 61.73% and 78.13 for BRAD and LARB datasets, respectively. The noteworthy performance of the suggested T-TC-INT text classification system across different sentence lengths serves as a testament to the effectiveness of employing the self-attention (SE-A) approach and Inductive Transfer (INT) framework and employing word units as input features to enhance the encoder's SE-A sub-layer proficiency in capturing word relationships within AD input sentences.

Table 11. Accuracy score on HARD dataset with different sentence lengths.

Sentence Length	Accuracy
<10	76.88%
(10–20)	77.02%
(20–30)	77.56%
(30–40)	79.03%
(40–50)	78.73%
>50	81.83%

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

Sentence Length	Accuracy
<10	50.45%
(10–20)	54.89%
(20–30)	57.48%
(30–40)	60.09%
(40–50)	60.87%
>50	61.73%

Table 13. Accuracy score on LARB dataset with different sentence lengths.

Sentence Length	Accuracy
<10	74.98%
(10–20)	75.37%
(20–30)	76.01%
(30–40)	77.09%
(40–50)	78.10%
>50	78.13%

4.7. Key Findings

- T-TC-INT Model for AD Classification: This research introduces a Transformer text classification model combined with an Inductive Transfer (INT) framework and with a self-attention (Se-F) approach for the classification of Arabic Dialects (ADs) into five-point categories. This architecture utilizes SE-A to enhance the representation of the global text sequence.
- Selective Term and Word Extraction: The SE-A approach employed in the model demonstrates the capability to select the most meaningful terms and words from the text sequences. This selective attention mechanism enhances the model's ability to capture essential information from the input.
- Quality Enhancement via Inductive Transfer (INT) and SE-A: Combining the benefits of the INT framework and using word units as input characteristics to the SE-A sub-

layer proves significant, especially for low-resource language text classification tasks, such as Arabic Dialects (ADs).

- Experimentation and Configuration Impact: Various experiments were conducted using different configurations, including the use of multiple heads in the SE-A sub-layer and training with multiple encoders. These experiments positively impacted the classification performance of the proposed system.
- Alternate Learning Outperforms Joint Learning: The findings reveal that alternate learning, as opposed to joint learning, yields better efficiency.
- Effect of Input Sentence Length: The effectiveness of the proposed T-TC-INT model increased with longer input sentence lengths, particularly for sentences with 30- to 40-word tokens and larger than 50-word tokens, achieving accuracy scores of 79.03% and 81.83%, respectively.
- State-of-the-Art Enhancement: The proposed model's practical experiment results showcase its superiority over existing approaches, evidenced by total accuracy percentages of 81.83% on the HARD dataset, 61.73% on the BRAD dataset, and 78.13% on the LARB dataset. This includes an improvement over well-known models like AraBERT and LR.

5. Conclusions

We introduce a T-TC-INT model tailored for the five-point categorization of ADs. The suggested design incorporates an SE-A methodology and employs the Inductive Transfer (INT) framework to enhance the comprehensive representation of the global text sequence. Moreover, the SE-A approach effectively identifies the most pertinent terms and phrases from within the text sequences. Through training on SA tasks for ADs, including both ternary and five-polarity tasks, the proposed system's effectiveness was notably bolstered. Leveraging the advantages of (SE-A) and Inductive Transfer (INT) significantly elevates the quality of the recommended text classification system. The outcomes of this study underscore the vital attributes of the T-TC-INT model, which leverages the SE-A approach to enhance accuracy across five-point and three-point classification tasks. The incorporation of the Inductive Transfer (INT) framework and word-unit characteristics into the SE-A sub-layer indicates their pivotal role in handling low-resource language text classification tasks, such as those in ADs. Similarly, conducting training with diverse configurations, such as employing multiple heads in the SE-A sub-layer and employing multiple encoders, heightened the classification performance of the suggested system. We conducted a series of experiments on two datasets for five-point Arabic SA. The results highlight that alternate learning methodologies yield superior efficacy compared to joint learning, a phenomenon influenced by the dataset sizes of each respective task. Furthermore, the findings demonstrate that the proposed system outperformed other cutting-edge techniques for the HARD, BRAD, and LARB datasets. Our investigation revealed that the performance of five-point classification could be enhanced by alternately approaching the tasks of fine-grained ternary classification within the Inductive Transfer (INT)-based suggested model. By identifying text as negative in the ternary setup, the distinction between high negative and negative categories in five-point classification can be refined. The empirical findings from the practical experiments, spanning both five-point and three-point classification tasks, conclusively demonstrated the enhanced accuracy of the suggested system when compared to alternative ADs text classification systems. The proposed T-TC-INT system excels in generating a robust latent feature representation for ADs text sequences. With an overall accuracy of 81.83% on HARD, 61.73% on BRAD, and 78.13% on LARB datasets, the experimental results underscore the superiority of the proposed T-TC-INT model over established state-of-the-art approaches like AraBERT [51], SVM [23], MNP [22], HC(KNN) [24], HC (KNN) [25], and LR [26]. It is worth noting that the proposed T-TC-INT system did not exhibit a significant improvement for the BRAD dataset in comparison to existing models. This discrepancy may be attributed to the fact that the BRAD Arabic dataset encompasses distinct domain-specific nuances, tones, and linguistic styles that

are not adequately captured by the proposed T-TC-INT Sentiment Analysis model. The absence of domain adaptation may result in a misalignment between the model's learned features and the unique characteristics of the BRAD dataset. The in-depth investigation of the experimental setup and outcomes revealed that the system's efficiency hinged on the utilization of the SE-A strategy and the dimensionality of word embedding. It was elucidated that incorporating the SE-A technique is advantageous, as it enables the extraction of both global and local semantic knowledge within the context by leveraging the SE-A sub-layer within each encoding layer. Additionally, the proposed T-TC-INT system effectively addresses the challenge of scarce training data in ADs. Furthermore, it successfully tackles the syntactic variability present in ADs phrases. Looking ahead, future endeavors will involve the development of an Inductive Transfer text classification framework incorporating sub-word units as input features to the SE-A sub-layer [62], as well as the implementation of a novel positional encoding mechanism [63] to address the syntactic and semantic intricacies inherent in right-to-left texts like ADs.

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

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

Data Availability Statement: The dataset generated during the current study is available in the [T-_TC_INT] repository (https://github.come/laith85, accessed on 1 November 2023).

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

References

1. Baniata, L.H.; Park, S.; Park, S.-B. A multitask-based neural machine translation model with part-of-speech tags integration for Arabic dialects. *Appl. Sci.* **2018**, *8*, 2502. [CrossRef]
2. Alali, M.; Mohd Sharef, N.; Azmi Murad, M.A.; Hamdan, H.; Husin, N.A. Multitasking Learning Model Based on Hierarchical Attention Network for Arabic Sentiment Analysis Classification. *Electronics* **2022**, *11*, 1193. [CrossRef]
3. Salloum, S.A.; AlHamad, A.Q.; Al-Emran, M.; Shaalan, K. A survey of Arabic text classification. *Inter. J. Elctr. Comput. Engi.* **2018**, *8*, 4352–4355.
4. Harrat, S.; Meftouh, K.; Smaili, K. Machine translation for Arabic dialects (survey). *Inf. Process. Manag.* **2019**, *56*, 262–273. [CrossRef]
5. El-Masri, M.; Altrabsheh, N.; Mansour, H. Successes and challenges of Arabic sentiment analysis research: A literature review. *Soc. Netw. Anal. Min.* **2017**, *7*, 54. [CrossRef]
6. Elnagar, A.; Yagi, S.M.; Nassif, A.B.; Shahin, I.; Salloum, S.A. Systematic Literature Review of Dialectal Arabic: Identification and Detection. *IEEE Access.* **2021**, *9*, 31010–31042. [CrossRef]
7. Abdul-Mageed, M. Modeling Arabic subjectivity and sentiment in lexical space. *Info. Process. Manag.* **2019**, *56*, 308–319. [CrossRef]
8. Al-Smadi, M.; Al-Ayyoub, M.; Jararweh, Y.; Qawasmeh, O. Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' reviews using morphological, syntactic and semantic features. *Info. Process. Manag.* **2019**, *56*, 308–319. [CrossRef]
9. Baly, R.; Badaro, G.; El-Khoury, G.; Moukalled, R.; Aoun, R.; Hajj, H.; El-Hajj, W.; Habash, N.; Shaban, K.; Diab, M. 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.; Babaoglu, 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. *Stud. Comput. Intell.* **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. *Pocedia Comput. Sci.* **2017**, *109*, 359–366. [CrossRef]
15. Alali, M.; Sharef, N.M.; Hamdan, H.; Murad, M.A.A.; Husin, N.A. Multi-layers convolutional neural network for twitter sentiment ordinal scale classification. *Adv. Intell. Syst. Comput.* **2018**, *700*, 446–454.
16. Alali, M.; Sharef, N.M.; Murad, M.A.A.; Hamdan, H.; Husin, N.A. Narrow Convolutional Neural Network for Arabic Dialects Polarity Classification. *IEEE Access* **2019**, *7*, 96272–96283. [CrossRef]
17. Gridach, M.; Haddad, H.; Mulki, H. Empirical evaluation of word representations on Arabic sentiment analysis. *Commun. Comput. Inf. Sci.* **2018**, *782*, 147–158.
18. Al Omari, M.; Al-Hajj, M.; Sabra, A.; Hammami, N. Hybrid CNNs-LSTM Deep Analyzer for Arabic Opinion Mining. In Proceedings of the 2019 6th International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019; pp. 364–368.
19. 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.
20. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 1243–1252.
21. Jin, N.; Wu, J.; Ma, X.; Yan, K.; Mo, Y. Inductive Transfer model based on multi-scale cnn and lstm for sentiment classification. *IEEE Access* **2020**, *8*, 77060–77072. [CrossRef]
22. Aly, M.; Atiya, A. LABR: A large scale Arabic book reviews dataset. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, 4–9 August 2013; Volume 2, pp. 494–498.
23. Al Shboul, B.; Al-Ayyoub, M.; Jararweh, Y. Multi-way sentiment classification of Arabic reviews. In Proceedings of the 2015 6th International Conference on Information and Communication Systems (ICICS), Amman, Jordan, 7–9 April 2015; pp. 206–211.
24. Al-Ayyoub, M.; Nuseir, A.; Kanaan, G.; Al-Shalabi, R. Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 531–539.
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. Inductive Transfer for Fine-Grained Twitter Sentiment Analysis. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, 7–11 August 2017; pp. 1005–1008.
29. Lu, G.; Zhao, X.; Yin, J.; Yang, W.; Li, B. Inductive Transfer using variational auto-encoder for sentiment classification. *Pattern Recognit. Lett.* **2020**, *132*, 115–122. [CrossRef]
30. Sohangir, S.; Wang, D.; Pomeranets, A.; Khoshgoftaar, T.M. Big Data: Deep Learning for financial sentiment analysis. *J. Big Data* **2018**, *5*, 3. [CrossRef]
31. Jangid, H.; Singhal, S.; Shah, R.R.; Zimmermann, R. Aspect-Based Financial Sentiment Analysis using Deep Learning. In Proceedings of the Companion of the Web Conference 2018 on The Web Conference, Lyon, France, 23–27 April 2018; pp. 1961–1966.
32. Ain, Q.T.; Ali, M.; Riaz, A.; Noureen, A.; Kamran, M.; Hayat, B.; Rehman, A. Sentiment analysis using deep learning techniques: A review. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 424.
33. Gao, Y.; Rong, W.; Shen, Y.; Xiong, Z. Convolutional neural network based sentiment analysis using Adaboost combination. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 1333–1338.
34. Hassan, A.; Mahmood, A. Deep learning approach for sentiment analysis of short texts. In Proceedings of the Third International Conference on Control, Automation and Robotics (ICCAR), Nagoya, Japan, 24–26 April 2017; pp. 705–710.
35. Qian, J.; Niu, Z.; Shi, C. Sentiment Analysis Model on Weather Related Tweets with Deep Neural Network. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Macau, China, 26–28 February 2018; pp. 31–35.
36. Pham, D.-H.; Le, A.-C. Learning multiple layers of knowledge representation for aspect based sentiment analysis. *Data Knowl. Eng.* **2018**, *114*, 26–39. [CrossRef]
37. Preethi, G.; Krishna, P.V.; Obaidat, M.S.; Saritha, V.; Yenduri, S. Application of deep learning to sentiment analysis for recommender system on cloud. In Proceedings of the 2017 International Conference on Computer, Information and Telecommunication Systems (CITS), Dalian, China, 21–23 July 2017; pp. 93–97.
38. Roshanfekr, B.; Khadivi, S.; Rahmati, M. Sentiment analysis using deep learning on Persian texts. In Proceedings of the 2017 Iranian Conference on Electrical Engineering (ICEE), Tehran, Iran, 2–4 May 2017; pp. 1503–1508.
39. Alharbi, A.S.M.; de Doncker, E. Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. *Cogn. Syst. Res.* **2019**, *54*, 50–61. [CrossRef]
40. Abid, F.; Alam, M.; Yasir, M.; Li, C.J. Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter. *Future Gener. Comput. Syst.* **2019**, *95*, 292–308. [CrossRef]

41. Vateekul, P.; Koomsubha, T. A study of sentiment analysis using deep learning techniques on Thai Twitter data. In Proceedings of the 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, 13–15 July 2016; pp. 1–6.
42. Pandey, A.C.; Rajpoot, D.S.; Saraswat, M. Twitter sentiment analysis using hybrid cuckoo search method. *Inf. Process. Manag.* **2017**, *53*, 764–779. [CrossRef]
43. Paredes-Valverde, M.A.; Colomo-Palacios, R.; Salas-Zárate, M.D.P.; Valencia-García, R. Sentiment analysis in Spanish for improvement of products and services: A deep learning approach. *Sci. Program.* **2017**, *2017*, 1329281. [CrossRef]
44. Patil, H.; Sharma, S.; Bhatt, D.P. Hybrid approach to SVM algorithm for sentiment analysis of tweets. In Proceedings of the AIP Conference, Virtual, 1 June 2023; Volume 2699. No. 1.
45. Luvembe, A.M.; Li, W.; Li, S.; Liu, F.; Xu, G. Dual emotion based fake news detection: A deep attention-weight update approach. *Inform. Proces. Manag.* **2023**, *60*, 103354. [CrossRef]
46. Lei, Y.; Yang, S.; Wang, X.; Xie, L. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transac Audio Speech Lang. Process* **2022**, *30*, 853–864. [CrossRef]
47. Li, W.; Li, Y.; Liu, W.; Wang, C. An influence maximization method based on crowd emotion under an emotion-based attribute social network. *Inf. Process. Manag.* **2022**, *59*, 102818. [CrossRef]
48. Vyas, P.; Reisslein, M.; Rimal, B.P.; Vyas, G.; Basyal, G.P.; Muzumdar, P. Automated classification of societal sentiments on Twitter with machine learning. *IEEE Trans. Technol. Soc.* **2022**, *3*, 100–110. [CrossRef]
49. Qureshi, M.A.; Asif, M.; Hassan, M.F.; Abid, A.; Kamal, A.; Safdar, S.; Akbar, R. Sentiment analysis of reviews in natural language: Roman Urdu as a case study. *IEEE Access* **2022**, *10*, 24945–24954. [CrossRef]
50. 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.
51. 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.
52. 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.
53. Al-Sabahi, K.; Zuping, Z.; Nadher, M. A hierarchical structured self attentive model for extractive document summarization (HSSAS). *IEEE Access* **2018**, *6*, 24205–24212. [CrossRef]
54. Dean, J.; Monga, R. TensorFlow. Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2015. Available online: https://www.tensorflow.org/ (accessed on 1 November 2023).
55. Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd.: Birmingham, UK, 2017.
56. Varoquaux, G.; Buitinck, L.; Louppe, G.; Grisel, O.; Pedregosa, F.; Mueller, A. Scikit-learn: Machine Learning in Python. *GetMobile Mob. Comput. Commun.* **2015**, *19*, 29–33. [CrossRef]
57. 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.
58. Pang, B.; Lee, L. *Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval*; Now Publishers: Boston, MA, USA, 2008; pp. 1–135.
59. Liu, S.; Johns, E.; Davison, A.J. End-to-end Inductive Transfer 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.
60. Baniata, L.H.; Park, S.; Park, S.-B. A Neural Machine Translation Model for Arabic Dialects That Utilizes multitask learning (MTL). *Comput. Intell. Neurosci.* **2018**, *2018*, 7534712. [CrossRef]
61. 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.
62. Baniata, L.H.; Ampomah, I.K.E.; Park, S. A Transformer-Based Neural Machine Translation Model for Arabic Dialects that Utilizes Subword Units. *Sensors* **2021**, *21*, 6509. [CrossRef]
63. Baniata, L.H.; Kang, S.; Ampomah, I.K.E. A Reverse Positional Encoding Multi-Head Attention-Based Neural Machine Translation Model for Arabic Dialects. *Mathematics* **2022**, *10*, 3666. [CrossRef]