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

Laith H. Baniata * and Sangwoo Kang *

Posted Date: 2 November 2023

doi: 10.20944/preprints202311.0187.v1

Keywords: Switch-Transformer; Mixture of Experts (MoE) Mechanism; Sentiment Analysis (SA); Arabic Dialects; 5-polarity, MTL.

Article

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

Laith H. Baniata [1,*] and Sangwoo Kang [1,*]

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

Abstract: In recent times, models like the Transformer have showcased remarkable prowess in tasks related to natural language processing. However, these models tend to be excessively intricate and demand extensive training. Additionally, while the multi-head self-attention mechanism in the Transformer model aims to capture semantic connections between words in a sequence, it encounters limitations when handling short sequences, thereby limiting its effectiveness in 5-polarity Arabic sentiment analysis tasks. The switch-transformer model has recently emerged as a high-performing alternative. Nevertheless, when these models are trained using single-task learning, they often fall short of achieving exceptional performance and struggle to generate robust latent feature representations, especially when working with compact datasets. This challenge is particularly pronounced in the case of the Arabic dialect, which is considered a low-resource language. Given these constraints, this research introduces a novel approach to sentiment analysis in Arabic text. This method leverages multitask learning in tandem with the switch-transformer shared encoder to enhance model adaptability and refine sentence representation. By introducing a mixture of expert (MoE) mechanism that break down the problem into smaller, more manageable sub-problems, the model becomes adept at handling lengthy sequences and intricate input-output relationships, benefiting both five-point and three-polarity Arabic sentiment analysis tasks. This proposed model effectively discerns sentiment in Arabic dialect sentences. The empirical results highlight the outstanding performance of the suggested model, as evidenced in evaluations on the Hotel Arabic-Reviews Dataset, the Book Reviews Arabic Dataset, and the LARB dataset.

Keywords: switch-transformer; mixture of experts (MoE) mechanism ; sentiment analysis (SA); arabic dialects; 5-polarity; MTL

MSC: 68T07

1. Introduction

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

 doi:10.20944/preprints202311.0187.v1

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

Moreover, an additional syntactic challenge in Arabic dialects (ADs) pertains to word arrangement. To dissect this matter, it is imperative to discern the positioning of the verb, subject, and object within an AD sentence. As previously outlined in the literature review, languages fall into distinct categories like subject-object-verb, exemplified by Korean, subject-verb-object, as seen in English, and verb-object-subject, as is the case with Arabic. Additionally, there are languages that allow for flexible word order, characteristic of ADs. Within AD expressions, this flexibility imparts advanced insights about the subject, object, and various other forms of information. Consequently, employing a single-task learning approach and relying solely on manually designed features prove inadequate for conducting sentiment analysis of Arabic dialects. Furthermore, these divergences within ADs present a formidable obstacle for conventional deep learning algorithms . This is because, with lengthier phrases in ADs, there is an influx of intricate and perplexing contextual details concerning the object, verb, and subject. A drawback of conventional deep learning methodologies is the loss of input sequence data, which leads to diminished performance of the sentiment analysis (SA) model as the input sequence lengthens. Additionally, the configuration of Arabic words' roots and characters can vary significantly depending on the context, as exemplified by (كتاب, كتابات, يكتب). Moreover, the absence of standardized orthographic conventions stands out as a primary challenge in ADs. This encompasses morphological distinctions across dialects, evident in the utilization of prefixes and suffixes absent in Modern Standard Arabic (MSA). Furthermore, it's worth noting that numerous Arabic words can convey multiple meanings based on the application of diacritics for the same syntax. Additionally, the development of deep-learning-powered sentiment analysis (SA) models necessitates a substantial corpus of training data, a resource that proves challenging to amass for Arabic dialects (ADs). These dialects are recognized as unstructured and resource-scarce languages, rendering the retrieval of information a formidable endeavor [1]. As the volume of training data diminishes for ADs, so does the classification efficacy. Moreover, the majority of Modern Standard Arabic (MSA) tools do not take into account the idiosyncrasies of Arabic dialects [2]. It is also important to note that relying solely on lexical resources like lexicons may not be the most effective approach for Arabic SA due to the vast array of words stemming from diverse dialects, making it improbable for any lexicon to encompass them all [3]. Furthermore, the creation of tools and resources tailored to Arabic dialects is a laborious and time-intensive undertaking [4].

In recent times, there has been a surge of research efforts focused on analyzing sentiments in Arabic dialects. This research is primarily dedicated to the classification of reviews and tweets, aiming to determine binary and ternary polarities. The majority of these methodologies [5-10] rely on lexicons, custom-crafted traits, and tweet-specific attributes, which serve as inputs for Machine Learning (ML) algorithms. Conversely, alternative approaches adopt a rule-based strategy, like the utilization of lexicalization principles. This involves establishing and prioritizing a set of heuristic

 doi:10.20944/preprints202311.0187.v1

rules to effectively categorize tweets into negative or positive sentiments [11]. On a different note, Arabic sentiment ontology introduces sentiments with varying degrees of intensity to discern user attitudes and facilitate tweet classification. Deep learning techniques for sentiment analysis, including Recurrent Neural Networks (RNNs) [12], Convolutional Neural Networks (CNNs) [13-16], and recursive auto-encoders, have garnered significant attention due to their remarkable adaptability and resilience achieved through automated feature extraction. Notably, the recently developed switch-transformer model [17] outperforms conventional transformr models [18], recurrent neural network (RNN)-based models in various natural language processing (NLP) tasks, thereby capturing the interest of researchers in the field of deep learning.

This research paper proposed a switch-transformer sentiment analysis (ST-SA) model, that utilize a blend of expert (MoE) mechanisms that break down the problem into smaller, more manageable sub-problems. The proposed model becomes adept at handling lengthy sequences and intricate input-output relationships, benefiting both five-point and three-polarity Arabic sentiment analysis tasks. Despite previous efforts to address the challenges of ADs sentiment analysis (SA), the approach of multi-task learning (MTL) has emerged as a promising solution. MTL enriches comprehension capabilities, elevates encoder quality, and augments the significance of sentiment classification compared to a conventional single-task classifier. This is achieved by concurrently processing related tasks, leveraging a shared representation of text sequences [19]. An important advantage of MTL lies in its ability to adeptly leverage various resources for akin tasks. However, it's noteworthy that most existing approaches for SA in ADs are predominantly focused on binary and ternary classifications. In this study, we redirect our attention to the five-polarity ADs SA problem, an area that, to our knowledge, has received limited investigation. Notably, the utilization of a switch-transformer architecture in conjunction with MTL for ADs SA classification has not been explored in prior studies. Previous methodologies addressing this classification concern primarily relied on conventional transformer and Bi-LSTM techniques. In summary, our contributions can be outlined as follows:

- This research article introduces a pioneering switch-transformer model that integrates multi-task learning (MTL) for sentiment analysis (SA) in Arabic Dialects (ADs). The proposed ST-SA model is founded on the a mixture of experts (MoE) mechnisim was developed to breaks down the problem into smaller, more straightforward sub-problems, enabling the model to effectively handle extended sequences and intricate input-output connections.
- Furthermore, a multi-head attention (MHA) mechanism was devised to capitalize on the correlation between three and five polarities through the utilization of a shared switch-transformer encoder layer. We elucidate the process of sequentially and collectively learning two tasks (ternary and five classifications) within the MTL framework. This approach aims to refine the representation of ADs text for each task and expand the scope of captured features.
- This research paper studied the effect of training the proposed switch-transformer model with varying embedding dimensions for each token, diverse token values, different attention head numbers, varying filter sizes, a diverse number of experts, a range of batch sizes, and multiple dropout values.
- The proposed SA-ST model employed a multi-head attention (MHA) mechanism to evaluate the correlation strength between two words within a sentence. This has notably bolstered the relevance and importance of various Natural Language Processing tasks.

The subsequent sections of this paper are structured as follows: Section 2 provides an overview of the literature, Section 3 delves into a detailed explanation of the proposed model, Section 4 showcases the experimental results, and lastly, Section 5 encapsulates the conclusions drawn from this study.

2. Literature Review

The Research in Arabic Sentiment Analysis (SA) regarding tasks with five levels of polarity classification has received comparatively less attention compared to those focused on binary or ternary polarity classification tasks. Moreover, the majority of approaches addressing this task rely on traditional machine learning algorithms. For instance, techniques based on corpora and lexicons

 doi:10.20944/preprints202311.0187.v1

were examined using Bag of Words (BoW) characteristics along with a range of machine learning algorithms including passive aggressive (PA), support vector machine (SVM), logistic regression (LR), naive Bayes (NB), perceptron, and stochastic gradient descent (SGD) for analysis on Arabic Book Review [20]. In a similar vein, [21] investigated the influence of stemming and the balancing of BoW characteristics using multiple ML algorithms on the same dataset. They discovered that stemming led to a reduction in performance. In [22], a divide-and-conquer strategy was suggested to address tasks involving ordinal-scale classification. Their model adopted a hierarchical classifier (HC) structure, where the five labels were subdivided into smaller sub-problems. It was observed that the HC model outperformed a single classifier. Building upon this foundation, diverse hierarchical classifier architectures were put forth [23]. These structures were pitted against ML classifiers like SVM, KNN, NB, and DT. The experimental outcomes indicated that the hierarchical classifier enhanced performance. Nevertheless, it's worth noting that many of these structures exhibit a decrease in performance.

Another investigation [24] scrutinized various Machine Learning classifiers, comprising LR, SVM, and PA, utilizing n-gram attributes in the Book Reviews in the Arabic Dataset (BRAD). The findings unveiled that SVM and LR attained the most commendable performances. Correspondingly, [25] assessed multiple sentiment classifiers, encompassing AdaBoost, SVM, PA, random forest, and LR on the Hotel Arabic-Reviews Dataset (HARD). They observed that SVM and LR exhibited superior performances when incorporating n-gram features. These previously mentioned methodologies highlight a notable absence of deep learning strategies for the five polarity classification tasks in Arabic Sentiment Analysis (SA). Moreover, a majority of the approaches addressing these five polarity tasks are grounded in traditional ML algorithms that rely on the feature engineering procedure, which is recognized as time-consuming and arduous. Additionally, these strategies are founded on Single-Task Learning (STL) and lack the capacity to discern the interrelationship among different tasks (cross-task transfer) and model various polarities concurrently, including both five and three polarities.

Other investigations have turned to Multitask Learning (MTL) to tackle the challenge of five-point Sentiment Analysis (SA) classification tasks. For instance, [26] introduced a multi-task learning framework utilizing a Recurrent Neural Network (RNN) to concurrently address both five-point and ternary classification tasks. Their model incorporated Bidirectional Long Short-Term Memory (Bi-LSTM) and Multilayer Perceptron (MLP) layers. Additionally, they enriched features with tweet-specific elements like punctuation counts, elongated words, emoticons, and sentiment lexicons. Their findings indicated that jointly training SA classification tasks significantly boosted the efficacy of the five-polarity task. Similarly, [27] leveraged the interplay between five-polarity and binary sentiment classification tasks by concurrently training them. The proposed model featured an encoder (LSTM) and a decoder (variational auto-encoder) as shared components for both tasks. Empirical results showcased that the Multitask Learning (MTL) model enhanced performance on the five-polarity task. The concept of Adversarial Multitasking Learning (AMTL) was initially introduced in [28]. This model incorporated two LSTM layers as task-specific components and one LSTM layer shared across tasks. Furthermore, a Convolutional Neural Network (CNN) was integrated with the LSTM, and the outputs from both networks were concatenated with the shared layer output, forming the final latent sentence representation. The authors noted that the proposed Multitask Learning (MTL) model elevated the performance of five-polarity classification tasks and bolstered the quality of the encoder. Although the Multitask Learning (MTL) methodologies detailed above have seen application in English, there is a conspicuous dearth of multi-task learning and deep learning techniques applied to five-polarity Arabic SA. Existing studies focused on this task have predominantly relied on single-task learning with traditional machine learning algorithms. Consequently, there remains ample room to augment the effectiveness of current Arabic SA methods in addressing the five polarities, as it still stands at a relatively modest level.

Further investigations have employed deep learning techniques for Sentiment Analysis (SA) across diverse domains, including finance [29,30], movie critiques [31-33], weather-related tweets [34], reviews on travel platforms [35], and cloud service recommendation systems [36]. As detailed

 doi:10.20944/preprints202311.0187.v1

by authors in [34], these studies autonomously extracted textual attributes from various data sources. They transformed weather-related knowledge and user information into word embedding using the word2vec tool. This method has been widely adopted in multiple research endeavors [29,37]. Jeong et al. [48] identified prospects for product development by amalgamating topic modeling and SA insights derived from social media data generated by customers. This approach serves as a real-time monitoring tool to analyze evolving customer preferences in dynamic product environments. Numerous studies have harnessed polarity-based sentiment deep learning techniques for analyzing tweets [38,39,40,41]. Researchers elucidated how deep learning methodologies enhanced the precision of their individual sentiment analyses. While the majority of these deep learning models have been applied to English text, there are a few models tailored to tweets in languages such as Persian [37], Thai [40], and Spanish [42]. These researchers conducted tweet analysis using a variety of models for polarity-based SA, including DNN, CNN, hybrid models [41], and SVM [43].

A multitude of techniques have been proposed to discern fabricated news, with a significant focus on linguistic traits, often disregarding the potential of dual emotional characteristics. Luvembe et al. [44] put forth a skillful approach to pinpointing misleading information by capitalizing on dual emotional attributes. To achieve this, the authors utilize a layered bi-GRU structure to extract emotional traits and encapsulate them within a feature vector framework. Additionally, the researchers integrate a profound attention mechanism atop the Bi-GRU to augment the model's capacity to grasp pivotal dual-emotion nuances. By employing the extracted emotional attributes as input, the authors employ the Adaptive Genetic Weight Update Random Forest, which gradually singles out the most relevant dual features. This iterative process substantially fortifies the classifier's precision in detection. To evaluate the proposed methodology, the authors conduct thorough experiments using three openly accessible datasets. The empirical results illustrate that the suggested model outperforms established techniques in efficiently recognizing fake news. In summary, the investigators introduce an inventive approach that incorporates dual emotion attributes, and the results affirm its superiority over customary approaches in the domain of counterfeit news identification.

Numerous scholarly articles have increasingly turned their attention to Twitter datasets as a foundational resource for constructing and refining sentiment analysis models. For instance, Vyas et al. [45] introduced an innovative hybrid framework that melds a lexicon-based approach with deep learning techniques to scrutinize and categorize the sentiments expressed in tweets related to COVID-19. The primary objective was to automatically discern the emotional tones conveyed in tweets concerning this subject. To accomplish this, the authors employed the VADER lexicon method to detect positive, negative, and neutral sentiments, which were then used to classify COVID-19-related tweets accordingly. During the categorization process, a range of machine learning (ML) and deep learning (DL) methods were deployed. Among these methods, the LSTM technique exhibited the most impressive performance, achieving a classification accuracy of 83%, outperforming the other approaches. Additionally, the ML classifier demonstrated a significant acceleration in processing speed, performing roughly ten times faster than the VADER technique. These findings underscore the potential of rapidly and automatically categorizing societal sentiments associated with COVID-19 on the Twitter platform. Such insights hold the potential to play a pivotal role in guiding public health awareness campaigns.

Baniata et al [46] introduce a novel approach utilizing a Multitask Learning Multi-Head Attention model for the five-point classification of ADs. This innovative architecture incorporates a Multi-Head Attention (MHA) technique and a Multitask Learning (MTL) framework to bolster the overall representation of text sequences. Moreover, the MHA method enables the selection of the most pertinent terms and phrases from these sequences. By training on Sentiment Analysis (SA) tasks, encompassing ternary and five-polarity tasks specific to ADs, the system's efficacy was significantly elevated. Leveraging the advantages of MHA and MTL amplifies the proficiency of the proposed SA system. The outcomes of this study underscore the pivotal attributes of the MTL-MHA SA system, which leverages the MHA method and heightens the accuracy of results for both five-point and three-point classification tasks. The incorporation of the MTL framework and word-units

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0187.v1

as input features for the MHA sub-layer indicates their critical role in low-resource language SA tasks, such as those involving ADs. Additionally, refining the model through diverse configurations, including employing multiple heads in the MHA sub-layer and training with multiple encoders, notably enhanced the classification performance of the suggested system. Furthermore, Alali et al. [47] introduced a multitasking methodology termed the Multitask Learning Hierarchical Attention Network (MTLHAN). This approach aims to augment the representation of sentences and enhance overall adaptability. The MTLHAN framework employs a shared word encoder and attention network for both tasks, utilizing two different training strategies to scrutinize three-polarity and five-polarity Arabic sentiment. The outcomes of the experiments emphasize the outstanding performance of this suggested model. Furthermore, when applied to the Arabic tweets dataset, the model displayed an exceedingly low macro mean absolute error of 0.632%, effectively surmounting the challenge of five-point Arabic sentiment classification.

3. The Proposed Switch-Transformer Sentiment Analysis Model that utilizes MoE Mechanism

Transformer-based models have exhibited remarkable efficacy across a spectrum of NLP tasks, encompassing the categorization of text. The conventional Transformer architecture [18], featuring multi-faceted self-focus, is a prevalent blueprint for this endeavor. As illustrated in Figure.1 ,its structure comprises an encoder composed of multiple tiers of multi-faceted self-focus and feedforward neural networks (FFN). This multi-faceted self-focus mechanism grants the model the capacity to assess the significance of various terms in a sequence grounded on their semantic associations, while the FFNs convert the output of the self-focus layer into a more advantageous representation. The crux of the Transformer is the self-focus mechanism founded on mathematical expressions [48]. Presented with a succession of input embedding $x_1,\ldots,x_n$ the self-focus mechanism derives a collection of contextually attuned embeddings $h_1,\ldots,h_n$ through the ensuing procedure :

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

where Attention is the scaled dot-product attention function:

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

Then, the multi-head attention is a concatenation of all head of h_i , as follows:

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

Additionally, the position-wise FFNs are multi-layer perceptrons applied independently to each position in the sequence, which provide a nonlinear transformation of the attention outputs. FNNs are calculated as follows:

$$FFN(x) = ReLU(xW_1 + b_1)\,W_2 + b_2 \qquad (4)$$

For each layer, there is a Layer Normalization which normalizes the inputs to a layer in a neural network to improve training speed and stability.

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0187.v1

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

By applying these procedures to every layer within both the encoder , the multi-head self-attention mechanism empowers the Transformer framework to encompass intricate semantic connections among words in a sequence, proving highly efficient across various natural language processing tasks. Nevertheless, the conventional Transformer design encounters specific limitations. A primary concern revolves around the self-attention (MHA) mechanism's quadratic computational demand concerning input sequence length, hindering scalability for exceptionally long sequences [49] and decreasing adaptability for shorter sequences . Furthermore, the self-attention mechanism (MHA) treats all positions in the input sequence uniformly, which might not be optimal for specific input types where certain positions hold greater significance than others. While the Transformer model has demonstrated outstanding performance in numerous NLP tasks, it may still grapple with capturing intricate input-output associations that necessitate more specialized models.

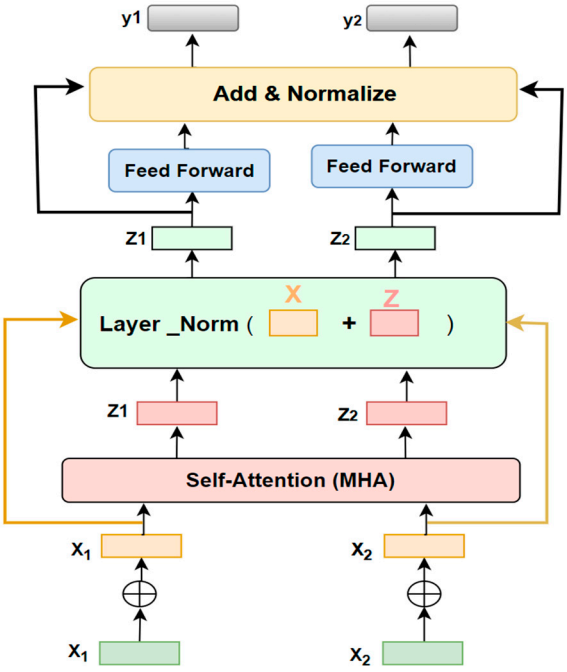

Figure 1. The Architecture of the Conventional Transformer Model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0187.v1

To surmount these obstacles, our research paper introduces a novel Switch-Transformer sentiment analysis (ST-SA) model, employing Multitask Learning (MTL) for the classification of Arabic Dialect (AD) sentiment. The objective behind adopting multi-task learning (MTL) is to augment the performance of the five-point Arabic sentiment analysis quandary, capitalizing on the interconnection between the ADs SA classification tasks, encompassing both five and ternary polarities. The proposed ST-SA model for ADs is based on the transformer model recently elucidated by Vaswani et al [18]. MTL has exhibited greater efficacy compared to singular-task learning. It harnesses the communal representation of diverse loss functions, concurrently handling SA tasks with three and five polarities, thereby refining the representation of both the semantic and syntactic facets of ADs text. The insights garnered from each task can fortify the learning process of other tasks, enhancing their efficacy. Furthermore, a pivotal facet of MTL lies in its provision of a superior approach to accessing resources devised for akin tasks, ultimately amplifying the learning proficiency of the current task and enriching the reservoir of exploitable knowledge. By means of comprehension, the layers involved in task-sharing can amplify the model's capacity for generalization, accelerate the pace of learning, and enhance its overall intelligibility. Similarly, leveraging the domain expertise embedded in the training cues of interconnected tasks as an inductive bias, the multi-task learning approach facilitates swift transfers that bolster generalization. This inductive transfer can be deployed to refine the precision of generalization, expedite the learning process, and heighten the transparency of the acquired models. A learner engaged in simultaneous acquisition of numerous interrelated tasks can employ these tasks as an inductive bias for one another, thereby gaining a more profound understanding of the domain's regularities. This can result in a more practical acquisition of Sentiment Analysis (SA) tasks for Arabic Dialects (ADs) even with a limited amount of training data. Similarly, multi-task learning collaboratively discerns the meaningful interrelation among the acquired tasks. As depicted in Figure.2, the proposed ST-SA sentiment analysis system boasts a distinctive architecture relying on multi-head attention (MHA) , MTL, a shared vocabulary, and specialized mechanisms referred to as a mixture of experts (MoE) inside the switching FNN layer . The presented ST-SA model, employing MTL, fine-tunes mixed classification tasks (ternary and five-polarity classification tasks) and comprehends them collectively. The integration of a shared switch-transformer block (encoding layer) streamlines the transfer of knowledge from the ternary task to the five-point task during the learning process, leading to an enhancement in the current task's (five-point task) learning capabilities.

Switch Transformers (ST) [17] aim to overcome the challenges of the trandiational Tranformer model by introducing a combination of expert mechanisms known as a mixture of experts (MoE). This strategy breaks down the problem into smaller, more straightforward sub-problems, enabling the model to effectively handle extended sequences and intricate input-output connections. As previously noted, while the multi-head self-attention mechanism in the Transformer model is designed to capture semantic ties between words in a sequence, it encounters limitations in dealing with concise sequences .The MoE mechanisms empower the model to partition the sequence into more manageable segments and apply distinct experts to each segment. This approach has led to enhancements in the model's performance on tasks involving shorter sequences and has attained leading-edge results in various benchmark evaluations [50], [51], [52]. The pivotal distinction in the mathematical formulation of the Switch Transformer in comparison to the traditional Transformer lies in the substitution of the Feed-Forward Network (FFN) with the Mixture of Experts (MoE) mechanism, as illustrated in Figure.3 . In the typical Transformer, the FFN is composed of two linear layers separated by a ReLU activation function. Conversely, the MoE mechanism employs a collection of expert networks to grasp distinct facets of the input data, subsequently amalgamating their outputs via a gating network. This enables the model to dynamically select from various parameter sets (expert modules) based on the input. This stands in contrast to the original Transformer model depicted in Equation 4, which employs a fixed set of parameters for all inputs. In a formal manner, the MoE mechanism within the Switch Transformer can be denoted by the subsequent equation:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0187.v1

The function $g_j(x_t)$ serves as a gate, influencing the significance of expert module j with respect to input x_t. Meanwhile, $e_j(x_t)$ represents the result produced by expert module j for input x_t. The switch mechanism operates by training the gating functions' parameters, which enable the dynamic selection of expert modules. This adaptive capability equips the model to accommodate diverse input patterns and excel across a range of tasks. The functioning of the MoE mechanism within the Switch Transformer has been done in several steps . First , the input undergoes partitioning into various subspaces, with each subspace undergoing individual processing by a distinct expert. Each of these experts constitutes an independent neural network that has been trained to excel in a particular subset of the input domain. Each expert generates an output vector that encapsulates its forecast for the specific input subspace provided. A gating process is employed to identify the expert most pertinent to a given input. This gating process takes the input and generates a series of weights that ascertain the significance of each expert's prediction. The final output is a weighted combination of the experts' forecasts, and the weighting for this amalgamation is dictated by the gating mechanism. In general, the MoE empowers the Switch Transformer to master intricate patterns within the input domain by capitalizing on the specialized expertise of numerous experts .This framework enables the model to glean insights from multiple experts, each adept in distinct facets of the data, and fuse their results to enhance overall performance. This can result in superior proficiency in tasks demanding thorough comprehension of inputs, presenting a hopeful remedy for the constraints of limited datasets in ADs text classification. Consequently, the study leverages this capacity to discern intricate relationships among words and phrases within ADs text.

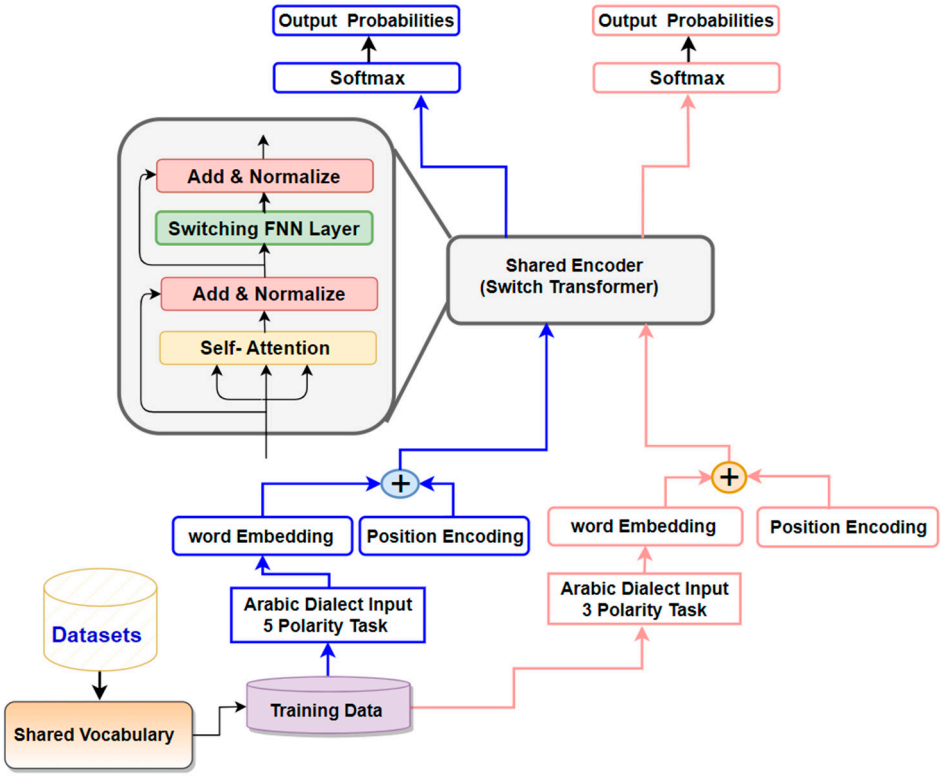

Figure 2. The Architecture of the Proposed Switch-Transformer Model for Arabic Dialects that utilizes MoE Mechanism.

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

4. Experiments:

 doi:10.20944/preprints202311.0187.v1

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

4.1. Data

The proposed model underwent training using three reference datasets. The initial dataset utilized was HARD [25], where reviews were collected from various reservation websites and categorized into five distinct groups. The second dataset, BRAD [24], and the third dataset, LARB [53], were subsequently employed for training purposes. The research project utilized review-level datasets, including BRAD, HARD, and LARB. BRAD's reviews were collected from the Goodreads website and categorized into five scales. The distribution of classes for HARD, BRAD, and LARB is detailed in Tables 1, 2, and 3, respectively. It's important to note that the datasets employed in this study were left in their unprocessed state, potentially affecting the reliability of the proposed model. Furthermore, all sentences underwent preprocessing, including the use of sentence breakers to segment reviews into individual sentences. Furthermore, any presence of Latin letters, non-Arabic characters, diacritics, hashtags, punctuation, and URLs were entirely removed from the texts of the ADs. The Arabic dialect texts underwent orthographic normalization for consistency. Additionally, emoticons were replaced with their corresponding descriptions, and adjustments were made for extended words. To prevent potential overfitting of the model, we implemented an early stopping technique with a patience parameter set to three epochs. When assessing the performance of the proposed ST-SA model that utilize MTL for sentiment analysis of Arabic vernaculars. We utilized a model checkpoint mechanism to save the most optimal weights of the proposed model. Besides being divided for training and testing purposes, the HARD, BRAD, and LARB datasets yield valuable insights into how polarities are distributed among their samples. The HARD dataset, with a total of 409,562 samples, is categorized into 5 polarities, each signifying distinct sentiments or attitudes. Allocating 80% of the dataset for training (327,649 samples) and 20% for testing (81,912 samples) ensures a comprehensive portrayal of the various polarities within both sets. Similarly, the BRAD dataset, comprising 510,598 samples, is divided with 80% (408,478 samples) earmarked for training and 20% (101,019 samples) for testing. Likewise, the LARB dataset, encompassing 63,257 samples, is split with 80% (50,606 samples) for training and 20% (12,651 samples) for testing. This partitioning approach guarantees that the 5-polarities are well-represented in both training and testing phases, allowing models to capture the subtleties of sentiment variation and effectively apply their understanding to unseen data. Prejudices can wield significant sway over the effectiveness of sentiment analysis models. If biases are present in the training data, they can skew the outcomes. To tackle this concern and determine the appropriate data selection for the presented ST-SA sentiment analysis model for Arabic vernaculars, we took into account five distinct steps:

- Guarantee that the training dataset comprises a multitude of origins and encompasses a broad spectrum of demographic profiles, geographic locales, and societal contexts. This approach serves the purpose of mitigating biases, resulting in a dataset that is not only more exhaustive but also more equitable in its composition.
- Confirm that the sentiment labels in the training dataset are evenly distributed among all demographic segments and viewpoints. This helps to reduce the risk of over-generalization and biases stemming from an unequal distribution of sentiment instances.
- Set forth precise labeling directives that explicitly guide human annotators to remain impartial and refrain from introducing their personal biases into the sentiment labels. This approach aids in upholding uniformity and reducing the potential for biases.
- Conducting an exhaustive examination of the training data to pinpoint potential biases is imperative. This entails scrutinizing factors like demographic disparities, serotype reinforcement, and any groups that may be inadequately represented. Upon identification, we implemented appropriate measures to rectify these biases. This involved employing techniques such as data augmentation, oversampling of underrepresented groups, and applying pre-processing methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0187.v1

Table 1. 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 2. 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 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. The Setup of the proposed model:

The introduced sentiment analysis model, known as Switch-Transformer that utilize Multitask Learning (ST-SA) , was created by harnessing the capabilities of TensorFlow [54], Keras [55], and scikit-learn [56] frameworks. To explore its effectiveness, a series of experiments were carried out for all ADs classification tasks, encompassing both three and five polarities. These experiments involved a diverse array of parameter configurations, specifically considering six different values for the word-embedding dimension of each token : 50, 32, 40, and 35. Additionally, the attention heads were assessed with six distinct values: 4, 2, 3. The position-wise FNN incorporated filters of varying dimensions, including 40, 30, 35, 32, and 50.

4.3. The Training Mechanisim of the proposed ST-SA model for Arabic Dialects:

Joint training and alternative training are two key approaches in the realm of multi-task learning models. Joint training involves training a model on multiple tasks simultaneously, sharing information and learning representations that are beneficial for all tasks. This approach leverages the interdependencies between tasks to improve overall performance. In contrast, alternative training focuses on training the model on tasks individually, cycling through them iteratively. This allows the proposed model to dedicate focused attention to each task, potentially yielding better performance for each task in isolation. Both approaches have their advantages and trade-offs. Joint training can lead to better generalization across tasks, while alternative training may excel in tasks with significant disparities in data distribution or complexity. The choice between these strategies hinges on the specific characteristics of the tasks at hand and the desired trade-offs in performance and efficiency. Ultimately, the selection of training methodology plays a crucial role in shaping the effectiveness and adaptability of multi-task learning models.

The proposed system adeptly handles both ternary and five-polarity classification tasks. For instance, during training on HARD, ST-SA alternates between instructing the model on the five-polarity and ternary classification tasks. We assessed two strategies for model training: alternating [57, 58] and joint learning. In our multi-task learning setup, we apply the loss and optimizer for each task in turns. This means that the training process kicks off with the ternary classification task for a specified number of epochs before transitioning to the five-polarity classification task. The primary aim in training both tasks was to minimize categorical cross-entropy. The suggested ST-SA model underwent training for 20 epochs, incorporating an early stopping mechanism set to activate after two epochs, with a batch size of 90. We adhered to standard protocols for the BRAD, HARD, and

 doi:10.20944/preprints202311.0187.v1

LARB datasets, splitting them into an 80% training subset and a 20% testing subset. We opted for the Aadam optimizer to instruct each task within the ST-SA model. We utilized a sentence breaker to segment reviews into sentences, with maximum sentence lengths set at 80 for BRAD, 50 for HARD, and 80 for LARB. The class weights methodology was not implemented in our proposed model [59]. Prior to each epoch, the training data undergoes a random shuffling process. Additional details on hyper-parameters can be found in section 4.5.

4.4. Sate-of-Art Approaches :

Employing the five-point datasets BRAD, HARD, and LARB for analyzing ADs, the ST-SA model designed for this purpose was assessed against the latest standard methods. Initially, Logistic Regression (LR) was introduced in [24] using unigrams, bi-grams, and TF-IDF, and was subsequently employed on the BRAD dataset. In a similar vein, Logistic Regression (LR) was initially advocated in [25] utilizing unigrams, bi-grams, and TF-IDF, and was then put into practice on the HARD dataset. The ST-SA model we propose has also been subjected to a comparative analysis using the LABR datasets. These reference methods encompass: SVM which employs a support vector machine classifier with n-gram characteristics, as recommended in [60]. MNB Implements a multinomial Naive Bayes approach with bag-of-words attributes, as outlined in [53]. HC which is a model utilizing hierarchical classifiers, constructed based on the divide-and-conquer technique introduced by [61] and HC(KNN) which is an enhanced iteration of the hierarchical classifiers model, still rooted in the divide-and-conquer strategy as delineated by [62]. In recent times, tasks in Natural Language Processing (NLP) achieved remarkable proficiency through the utilization of the bi-directional encoder representation from transformers, known as BERT [63]. The AraBERT [64], an Arabic pre-trained BERT model, underwent training on three distinct corpora: OSIAN [65], Arabic Wikipedia, and the MSA corpus, encompassing a staggering 1.5 billion words. We've conducted a comparative analysis between the proposed ST-SA system for ADs and AraBERT [64], which boasts 786 latent dimensions, 12 attention facets, and a composition of 12 encoder layers.

4.5. Results:

Numerous empirical experiments were conducted employing the proposed ST- SA system for Arabic dialects. The suggested ST- SA system underwent training with varying configurations of attention heads (AH) in the MHA sub-layer and diverse encoder quantities to ascertain the most efficient structure. Additionally, the system was trained with varying dimensions of word embeddings for each token. This research delved into the influence of training the proposed system using two multitasking methodologies, namely, in tandem and alternatively, for performance assessment. The efficacy of the suggested system's sentiment analysis was assessed using an automated accuracy metric. This section details the evaluation of the proposed ST- SA system across five polarity classification tasks for ADs. The results of the practical experiments on HARD , BRAD and LARB are delineated in Tables 4, 5, and 6, respectively. The efficiency of the suggested ST- SA model, under the application of joint and alternate learning methods to HARD and BRAD, is succinctly summarized in Table 10, respectively. As elucidated in Figure 4, Tables 4 and 7, the proposed ST- SA system achieved an accuracy of 84.02% on the HARD-imbalanced dataset, where the number of AH was 2, number of tokens was 90, number of experts was 10 ,batch_size was 60 ,filter size was 32 , dropout value was 0.25 ,and the embedding dimension for each token was 23. This commendable accuracy was attained due to the favorable impact of employing the MTL framework , MoE mechanism and MHA approach, particularly in right-to-left texts like ADs. MoE employs a collection of expert networks to grasp distinct facets of the input data, subsequently amalgamating their outputs via a gating network. This enables the model to dynamically select from various parameter sets (i.e., expert modules) based on the input amd the,so that the proposed model can detect the sentiments accuretly.When juxtaposed with the performance of the top-performing system on the HARD dataset, the outcomes produced by the ST- SA model surpassed those obtained by LR [66], exhibiting an accuracy differential of 7.92%. Moreover, the proposed model outshone AraBERT [64], with an accuracy differential of 3.17% and the proposed ST-SA model outperformed the MTL-

 doi:10.20944/preprints202311.0187.v1

MHA SA [46] with an accuracy differential of 2.19%. Consequently, the concurrent execution of learning-related tasks augmented the pool of usable data and mitigated the risk of overfitting [67]. The presented system demonstrated proficiency in capturing both syntactic and semantic attributes, enabling it to discern the sentiments conveyed in AD sentences.

Furthermore, the recommended ST- SA system demonstrated superior effectiveness on the imbalanced BRAD dataset. As depicted in Table 5, the proposed model achieved an accuracy of 68.81%, where the number of AH was 3, number of tokens was 24, number of experts was 15 ,batch_size was 53 ,filter size was 30 , dropout value was 0.24 ,and the embedding dimension for each token was 50. As elucidated in Table 8, the suggested ST-SA system surpassed the logistic regression (LR) approach advocated by [24], exhibiting an accuracy differential of 21.71%, and outperformed the AraBERT model [64] by a margin of 7.96%. Also , the proposed ST-SA surpassed the MTL-MHA [46] SA system with an accuarcy differential of 7.08% .Additionally, the integration of the Switch-Transformer-based shared encoder (one for each classification task) enabled the suggested model to glean a comprehensive representation, encompassing the preceding, subsequent, and localized contexts of any position within a sentence.

Moreover, the suggested Switch-Transfirmer Sentiment Analysis model that utilize Multitask Learning (ST- SA), detailed in Table 6, exhibited exceptional performance on the demanding LARB imbalanced dataset. In this investigation, this innovative model attained a noteworthy accuracy of 83.91%, surpassing alternative methodologies. It's worth noting that with a specific setup comprising three Attention Heads (AH), a filter size of 35, number of tokens of 100, number of experts of 12 ,batch_size set to 70, dropout value set to 0.27 ,and the embedding dimension for each token set to 60, the suggested system truly showcased its effectiveness. This accomplishment underscores the resilience of the ST- SA model in navigating the intricacies of sentiment analysis within the framework of an imbalanced dataset. As demonstrated in Table 9, the porposed Switch-Transfirmer Sentiment Analysis model that utilize Multitask Learning (ST-SA), system exhibited its superiority over several alternative approaches. Notably, the ST-SA model outperformed various models by substantial margins. For instance, it displayed a remarkable accuracy differential of 33.61% when compared to the SVM [60] model, an impressive 38.91% accuracy differential surpassing the MNP [53] model, a significant 26.11% accuracy differential over the HC(KNN) [61] model, as well as a noteworthy 24.95% accuracy differential compared to AraBERT [64] and the proposed model even surpassed HC(KNN) [62] by an accuracy differential of 11.27%. Additionally , the proposed model surpassed the MTL-MHA SA Model [46] with accuracy differential of 5.78%.

Joint training, within the realm of deep learning, involves the concurrent training of a single neural network model to undertake multiple interrelated tasks. Rather than training distinct models for each task, this approach enables the model to collaborate and learn shared representations that can be advantageous for all tasks. This can lead to enhanced adaptability, heightened efficiency, and potentially even superior performance on each specific task. Imbalanced data signifies an uneven distribution of classes (or categories) within a dataset. In certain instances, one or more classes may possess notably fewer instances in comparison to others. This situation can present difficulties for deep learning models as they might exhibit a bias towards the majority class, resulting in subpar performance on minority classes. The evaluation results suggest that the presented ST- SA system, when subjected to both joint and alternate learning, exhibited exceptional efficiency. Alternate training outperformed joint learning, yielding accuracies of 84.02% and 76.62% in the imbalanced HARD dataset, and 67.37% and 64.23% in BRAD, as outlined in Table 10. Upon comparison with benchmark methods, it became evident that alternate training in five-point classification can yield more comprehensive feature representations within the text sequence than a single learning task. These outcomes highlight that alternate learning is better suited for tackling complex SA tasks, and it can grasp and generate a more robust latent representation in intricate tasks for AD SA. The discernible contrast in effectiveness between the two methodologies lies in how alternate learning is influenced by the volume of data in each task's dataset. Shared layers tend to hold more information when a task encompasses a larger dataset. In contrast, joint learning may lean towards bias if one of the tasks is associated with a significantly larger dataset than the other. Consequently, alternative

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0187.v1

training methods are deemed more suitable for tasks related to sentiment analysis of Arabic dialects. This holds particularly true in scenarios where there are two distinct datasets for different tasks, such as in machine translation tasks where translation is conducted from AD to MSA and then to English [57]. The efficacy of each task can be augmented by designing a network in an alternate configuration, obviating the need for additional training data [58]. Moreover, related tasks can further bolster the efficiency of five-point classification. The significance of the enhancements observed in our proposed model's performance can be attributed to multiple factors. Surpassing state-of-the-art models like AraBERT and LR stands as a notable achievement in itself, considering AraBERT's established effectiveness in Arabic language processing tasks. By outperforming AraBERT on the same datasets, our proposed model demonstrates its heightened precision in handling Arabic dialects. Additionally, even slight enhancements in accuracy bear significance as they contribute to elevating the overall performance of models designed for processing Arabic dialects. These incremental improvements can have practical implications, including refining the accuracy of sentiment analysis, information retrieval, and other applications in natural language processing for Arabic dialects.

Table 4. Results for the prposed ST- SA model on HARD dataset for the **five-polarities** classification task.

E-D-T	NT	AH	FS	NE	BS	DO	Accuracy (5-Polarity)
50	50	4	50	10	50	0.30	81.39%
32	100	2	32	10	50	0.25	83.81%
23	**90**	**2**	**32**	**10**	**60**	**0.25**	**84.02%**
30	150	4	30	5	50	0.25	82.89%
30	25	4	30	5	50	0.30	82.72%

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.

Table 5. Results for the ST- SA model on BRAD dataset for the **five-polarities** classification task.

E-D-T	NT	AH	FS	NE	BS	DO	Accuracy (5-Polarity)
30	20	2	30	6	40	0.22	66.72%
40	15	3	30	10	55	0.25	67.37%
35	17	3	35	13	52	0.30	64.95%
50	**24**	**3**	**30**	**15**	**53**	**0.24**	**68.81%**
55	30	3	40	18	56	0.26	67.15%

Table 6. Results for the ST- SA model on LARB dataset for the **five-polarities** classification task.

E-D-T	NT	AH	FS	NE	BS	DO	Accuracy (5-Polarity)
40	20	3	35	10	50	0.30	80.09%
60	**100**	**3**	**35**	**12**	**70**	**0.27**	**83.91%**
35	40	2	40	10	60	0.20	81.74%
20	40	4	39	15	40	0.30	82.65%

Table 7. The performance of the proposed ST- SA model compared with benchmark approaches on HARD dataset.

Model	Polarity	Accuracy
LR [66]	5	76.1%
AraBERT [64]	5	80.85%
MTL-MHA-SA [46]	5	81.83%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0187.v1

The proposed ST-SA Model	**5**	**84.02%**

Table 8. The performance of the proposed ST- SA. model compared with bench mark approaches on BRAD dataset.

Model	Polarity	Accuracy
LR [24]	5	47.7%
AraBERT [64]	5	60.85%
The MTL-MHA SA [46]	5	61.73%
The proposed ST-SA Model	**5**	**68.81%**

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

Model	Polarity	Accuracy
SVM [60]	5	50.3%
MNP [53]	5	45.0%
HC(KNN) [61]	5	57.8%
AraBERT [64]	5	58.96%
HC(KNN) [62]	5	72.64%
MTL-MHA SA [46]	5	78.13%
The Proposed ST-SA Model	**5**	**83.91%**

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

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

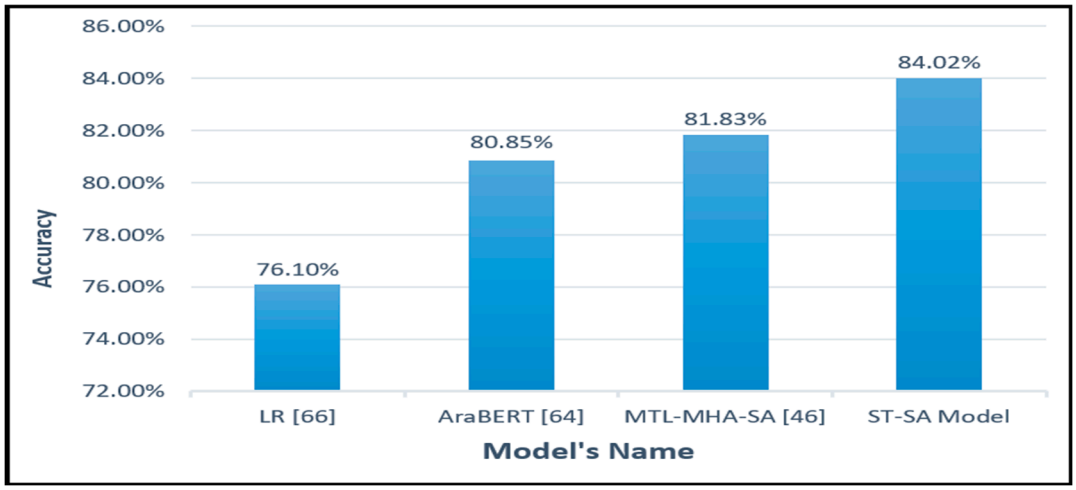

Figure 4. The Evaluation Accuracy of the Proposed ST-SA Model in Comparison with Sate-of-Art Approaches on HARD Test Dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2023 doi:10.20944/preprints202311.0187.v1

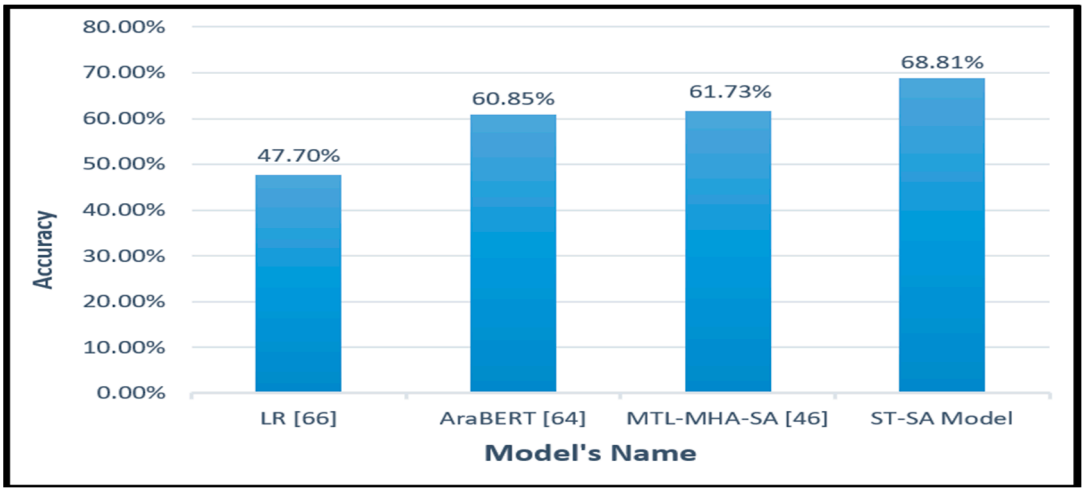

Figure 5. The Evaluation Accuracy of the Proposed ST-SA Model in Comparison with Sate-of-Art Approaches on BRAD Test Dataset.

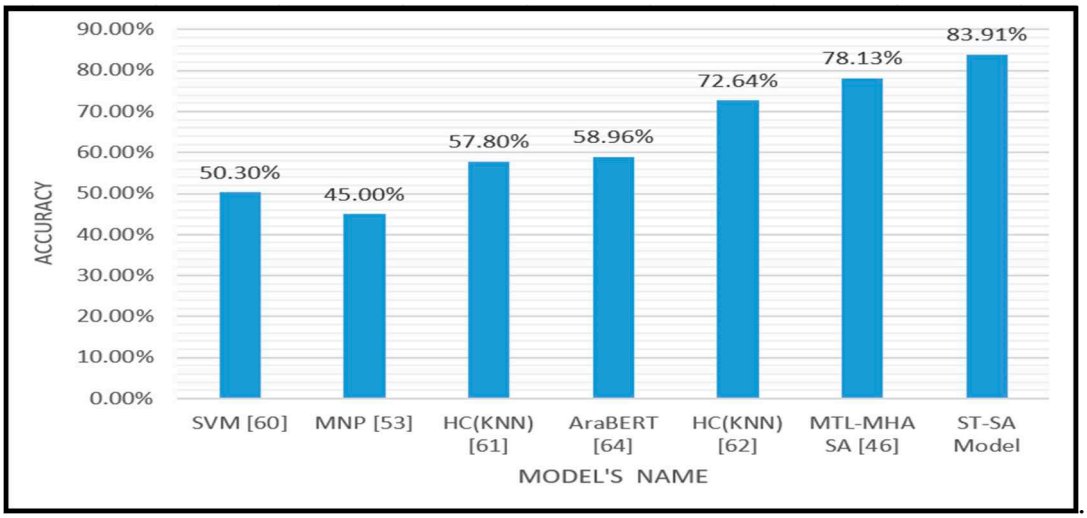

Figure 6. The Evaluation Accuracy of the Proposed ST-SA Model in Comparison with Sate-of-Art Approaches on LARB Test Dataset.

4.6. Impact of number of experts NE:

As demonstrated in Tables 4, 5, and 6, the effectiveness of the recommended ST-SA framework across diverse input representations derived from the self-attention layer underscores the significance of the proposed model for the classification task encompassing five distinct polarities. Here, "NE" denotes the number of experts in the encoding layer within the suggested switch-transformer SA model, employing the MoE mechanism. The devised system underwent training utilizing varying expert numbers: 5, 6, 10, 12, 13, 15, and 18. As evident from Tables 4, 5, and 6, a discernible shift in accuracy scores is observed for the categories HARD, BRAD, and LARB.

4.7. Impact of Length of Input Sentence:

Acquiring extended syntactic dependencies and contextual comprehension across elements in input expressions enhances the efficacy of classifying lengthy sentences. Sentences of equivalent length (in terms of source tokens) were clustered together as demonstrated in the work of Luong et al. [68]. Due to the substantial scale of the HARD corpus, a task involving a five-fold classification of polarities on the HARD dataset was selected to assess the performance of the Self-Attention (SA) mechanism for prolonged sentences. The assessment in this section is predicated on the subsequent

 doi:10.20944/preprints202311.0187.v1

ranges: <10, 10-20, 20-30, 30-40, 40-50, and >50. An automated accuracy measure was computed for the output generated by the Switch Transformer Sentiment Analysis system. As depicted in Table 11, the effectiveness of the recommended Switch-Transformer Sentiment analysis (ST- SA) model escalated with the elongation of input sentence length, particularly in the case of 40 to 50-word tokens and those surpassing 50-word tokens, registering accuracy scores of **81.32** and **84.02**, respectively. Through the employment of multi-task learning, a multi-headed attention mechanism, mixture of experts (MoE) mechanism ,and the incorporation of word units as an input characteristic for the MHA sub-layer, the proposed system attains contextually pertinent knowledge and dependencies within tokens, regardless of their position in the ADs input phrase. Furthermore, the utilization of MoE in the Switch Transformer enables it to excel at discerning complex patterns within the input domain, leveraging the specialized knowledge of a multitude of experts. However, the efficiency of the suggested model was notably lower for shorter sentences, specifically those comprising 10 to 20-word tokens , 30 to 40-word tokens, and 20 to 30-word tokens. Furthermore, the system's effectiveness notably dipped for sentences with fewer than 10-word tokens, yielding a meager accuracy of **77.25.** The impressive performance of the recommended ST- SA system across various sentence lengths underscores the efficacy of leveraging the MHA methodology and MTL framework, along with employing Mixture of Experts (MoE) mechanism, in enhancing the encoder's MHA sublayer proficiency in discerning word relationships within the ADs input sentence.

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

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

4.8. Principal Findings:

- The study proposes an innovative approach, the switch-transformer multi-task learning (ST-MTL) Model, for classifying Arabic Dialects (ADs) into five distinct categories. This method combines Multi-Task Learning (MTL) with a cutting-edge switch-transformer model. The incorporation switch-transformer and particularly the Mixture of Experts (MoE) mechanism serves to augment the portrayal of the comprehensive text sequence on a global scale.
- The model's capability to draw insights from various experts, each specializing in different aspects of the data, allows it to combine their findings, thereby boosting overall performance. This can lead to a heightened proficiency in tasks requiring a deep understanding of inputs, offering a promising solution to the limitations posed by restricted datasets in text classification for AD.
- Elevating Quality through MTL and switch-transfomer: The amalgamation of the Multi-Task Learning (MTL) framework with the incorporation of word-units as input attributes to the MHA sub-layer in switch-transformer encoder yields noteworthy advantages. This synergy is particularly pronounced in low-resource language Sentiment Analysis endeavors, exemplified by tasks involving Arabic Dialects.
- Superior Performance of Alternate Learning Over Joint Learning: The results indicate that opting for alternate learning, rather than joint learning, leads to enhanced effectiveness.
- Impact of Input Sentence Length: The efficacy of the suggested ST-SA model amplified as the length of input sentences extended, notably for sentences comprising 40 to 50-word tokens and those surpassing 50-word tokens, attaining impressive accuracy scores of **81.32**% and **84.02**%, respectively.
- Cutting-Edge Advancement: The empirical findings from the practical experimentation of the suggested model clearly demonstrate its supremacy over current methodologies. This is substantiated by the remarkable total accuracy rates of **84.02**% on the HARD dataset, **68.81**% on

 doi:10.20944/preprints202311.0187.v1

the BRAD dataset, and **83.91**% on the LARB dataset. Notably, this represents a notable enhancement compared to renowned models such as The MTL-MHA SA ,AraBERT and LR.

5. Conclusion:

We introduce an ST-SA model designed for the five-point categorization of Arabic Dialects (ADs). The suggested framework leverages the self-attention approach and a Multi-Task Learning (MTL) framework to enrich the global representation of the text sequence. Moreover, the MHA methodology is adept at singling out the most pertinent terms and words within the text sequences. Through training on Sentiment Analysis (SA) tasks encompassing ternary and five-polarity assignments for ADs, the system's effectiveness was notably enhanced. The utilization of MHA in conjunction with MTL markedly elevates the quality of the proposed SA system. The outcomes of this study underscore the pivotal attributes of the ST-SA system, which employs the MoE mechanism, MHA approach to augment accuracy in both five-point and three-point classification tasks. The integration of MTL framework, MoE mechanism ,and word units as input characteristics to the MHA sub-layer underscores the critical role of these strategies in low-resource language SA tasks, such as ADs. Similarly, experimenting with various configurations, including the deployment of multiple heads in the MHA sub-layer and training with multiple number of experts empowers the prposed ST-SA to master intricate patterns within the input domain by capitalizing on the specialized expertise of numerous experts. Also, it led to a notable boost in the classification performance of the proposed system. Conducting a series of experiments on two datasets for five-point Arabic SA, our findings reveal that alternate learning paradigms demonstrate superior efficiency compared to joint learning, with the dataset size of each task exerting an influence. The results unequivocally demonstrate that the proposed system outperforms other cutting-edge techniques across the HARD, BRAD, and LARB datasets. We further discerned that the five-point classification performance could be significantly enhanced by alternately addressing the tasks of fine-grained ternary classification within the MTL framework-based model. This fine-grained approach, particularly in designating text as negative in ternary configuration, contributes to a refined discrimination between the high negative and negative categories in the five-point classification schema.

The outcomes of practical experiments conducted on five-point and three-point categorization tasks have elucidated that the recommended system substantially enhanced accuracy when compared to other sentiment analysis systems for Arabic dialects. The proposed switch-Transformer Sentiment Analysis (ST-SA) system that’s utilize MTL generates a resilient latent feature representation for textual sequences in Arabic dialects. With overall accuracy rates of **84.02**%, **68.81**%, and **83.91**% on the HARD, BRAD, and LARB datasets correspondingly, the empirical findings underscore the superior performance of the ST-SA model over existing state-of-the-art methodologies, such as AraBERT [64], Support Vector Machine (SVM) [60], Multi-Neural Perceptron (MNP) [53], Hierarchical Clustering with K-Nearest Neighbors HC(KNN) [61], HC(KNN) [62], and Logistic Regression (LR) [24] and MTL-MHA SA [46]. Notably, the ST-SA system did not exhibit substantial improvements on the BRAD dataset when compared to existing models. This can be attributed to the BRAD Arabic dataset's domain-specific idiosyncrasies, tones, and linguistic styles, which the proposed ST-SA Sentiment Analysis model does not adequately capture. The absence of domain adaptation can result in a misalignment between the model's learned features and the distinctive characteristics of the BRAD dataset. Further analysis of the experiments and outcomes unveiled that the system's efficacy is contingent on the utilization of the Multi-Head Attention (MHA) strategy and the dimensionality of word embeddingsfor each token. The practical investigation elucidated the advantages of employing the MHA technique, as it enables the extraction of both global and local semantic knowledge within the contextual framework through the MHA sub-layer in each encoding layer. Additionally, the proposed ST-SA system addresses the challenge of limited training data and tackles the syntactic issue inherent in the free-format nature of AD phrases. Moreover, the ST-SA system, incorporating the MHA strategy,MoE meachnims , and word units as input features to the MHA sub-layer, demonstrated proficient sentence classification for ADs. Future plans encompass the development of a multi-task learning SA architecture utilizing sub-word units

as input features for the MHA sub-layer [69], as well as the adoption of a novel positional encoding mechanism and CNN approach [70] to effectively address the syntactic and semantic intricacies encountered in right-to-left textual content, such as ADs.

Author Contributions: L.H.B., S.K. conceived and designed the methodology and experiments; L.H.B. performed the experiments; L.H.B. analyzed the results; L.H.B., 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 [ST_SA_AD] repository (https://github.come/laith85).

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

References

1. Salloum, S.A.; AlHamad, A.Q.; Al-Emran, M.; Shaalan, K. A survey of Arabic text classification. *Inter Journal Elctre Comput Engi*. 2018,8,4352-4355.
2. Harrat, S.; Meftouh, K.; Smaili, K. Machine translation for Arabic dialects (survey). *Inf .Process. Manag* . 2019,56 ,262-273.
3. 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.
4. 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.
5. Abdul-Mageed, M. Modeling Arabic subjectivity and sentiment in lexical space. *info.process.Manag*.2019,56,308-319.
6. 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.
7. 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.
8. El-Beltagy, S.R.; El Kalamawy, M.; Soliman, A.B. NileTMRG at SemEval-2017 Task 4: Arabic Sentiment Analysis. In proceeding of the 11th International Workshop on Semantic Evaluation (semEval-2017), Vancouver, BC, Canada ,3-4 August 2017; pp.790-795.
9. 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,
10. Mulki, H.; Haddad, H.; Gridach, M.; Babao ˘glu, 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.
11. Siddiqui, S.; Monem, A. A; Shaalan, K. Evaluation and enrichment of Arabic sentiment analysis. *Stud.Compu. Intell*. 2017,740,17-34.
12. 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.
13. 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.
14. 14. 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.
15. Gridach, M.; Haddad, H.; Mulki, H. Empirical evaluation of word representations on Arabic sentiment analysis. *Commun. Comput. Inf. Sci*. 2018, 782, 147–158.
16. 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.
17. W. Fedus and et. al., "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," J. Mach. Learn. Res, vol. 23, pp. 1–40, 2021.
18. 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.

19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. Elnagar, A.; Khalifa, Y.S.; Einea, A. Hotel Arabic-reviews dataset construction for sentiment analysis applications. *Stud. Comput. Intell*. 2018, 740, 35–52.
26. 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.
27. 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
28. 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.
29. Sohangir, S.; Wang, D.; Pomeranets, A.; Khoshgoftaar, T.M. Big Data: Deep Learning for financial sentiment analysis. J. Big Data 2018, 5, 3.
30. Jangid, H.; Singhal, S.; Shah, R.R.; Zimmermann, R. Aspect-Based Financial Sentiment Analysis using Deep Learning. In Proceedings of the Companion of the The Web Conference 2018 on The Web Conference, Lyon, France, 23–27 April 2018; pp. 1961–1966.
31. 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.
32. 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.
33. 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.
34. 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.
35. Pham, D.-H.; Le, A.-C. Learning multiple layers of knowledge representation for aspect based sentiment analysis. Data Knowl. Eng. 2018, 114, 26–39.
36. 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.
37. 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.
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. 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.
41. Pandey, A.C.; Rajpoot, D.S.; Saraswat, M. Twitter sentiment analysis using hybrid cuckoo search method. Inf. Process. Manag. 2017, 53, 764–779.
42. 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.
43. Patil, H.; Sharma, S.; Bhatt, D. P. Hybrid approach to SVM algorithm for sentiment analysis of tweets. In Proceedings pf AIP conference , June 2023; Vol. 2699, No. 1.

44. 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(4), 103354.
45. 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 Transac on Tech and Society* **2022**, *3*(2), 100-110.
46. Baniata, L.H.; Kang, S. Multi-Head Attention Based-Sentiment Analysis Model for Arabic Dialects that utilizes Shared Vocabulary. Electronics 2023
47. 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.
48. T. Lin and et. al., "A survey of transformers," AI Open, 2022
49. C. Raffel and et. al., "Exploring the limits of transfer learning with a unified text-to-text transformer," The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 2020.
50. F. Xue and et. al., "Go wider instead of deeper," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 8, 2022, pp. 8779–8787.
51. A. Lazaridou and et. al., "Mind the gap: Assessing temporal generalization in neural language models," Advances in Neural Information Processing Systems, vol. 34, pp. 29 348–29 363, 2021.
52. A. Fan and et. al., "Beyond english-centric multilingual machine translation," The Journal of Machine Learning Research, vol. 22, no. 1, pp. 4839–4886, 2021.
53. 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.
54. Dean, Jeff, and Rajat Monga'TensorFlow. "Large-Scale Machine Learning on Heterogeneous Distributed Systems'." TensorFlow. org (2015).
55. Gulli A, Pal S. Deep learning with Keras. Packt Publishing Ltd; 2017 Apr 26.
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.
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. Pang, B.; Lee, L. Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval; Now Publishers: Boston, MA, USA, 2008; pp. 1–135.
67. 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.
68. 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.
69. 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.

 doi:10.20944/preprints202311.0187.v1

70. 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.