Journal of Theoretical and Applied Information Technology
15th May 2022. Vol.100. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# MODELING TEXT GENERATION WITH CONTEXTUAL FEATURE REPRESENTATION AND DIMENSIONALITREDUCTION USING DEEP TRANSFER LEARNING AND BI-LSTM

**K YOGESWARA RAO[1] , K SRINIVASA RAO[2]**

[1]Assistant Professor, GITAM (Deemed to be University) , Department of Computer Science and Engineering, India

[2]Associate Professor, GITAM (Deemed to be University) , Department of Computer Science and Engineering, India

E-mail: [1]yogiindsisu@gmail.com, [2]srinivas6591@gmail.com

**ABSTRACT**

The text generation models encounter challenges in learning the high-dimensional and abundance of data from the source domain, which leads to inaccurate text generation. In addition, generating the text sequence with contextual knowledge is challenging, even training the learning model with the domain knowledge. Hence, it is essential to effectively extract the knowledge from the source domain in the transfer learning model, which necessitates the contextual embedding of the input sequence and dimensionality reduction. Thus, this work develops the Deep Transfer Learning-based Text Generation (DTGEN) with the Embedding from Language Model (ELMo), Variational AutoEncoder (VAE), and Bidirectional-Long Short Term Memory (Bi-LSTM). The proposed DTGEN model focuses on the low-dimensional text feature representation and the deep transfer learning-based text generation. Initially, the proposed DTGEN approach contextually models the input text sequence with the help of the ELMo and generates the vectorized input data as the feature representation. Moreover, it generates the low-dimensional feature vector for the embedding input-based knowledge transfer by adopting the VAE model with the weighted regularization. In subsequence, the DTGEN approach learns the relevant knowledge from the source domain based on the contextual vector representation and analyzes the sequential patterns of the text data in the target domain. Consequently, the proposed approach predicts the terms based on the sequential relationships in the target text and the knowledge extracted from the source text with the assistance of the Bi-LSTM model. Thus, the experimental results demonstrate that the proposed DTGEN approach yields 2.89% higher text generation accuracy than the existing Deep Learning-based Next Word Prediction (DNWP) approach while testing on the NewsRoom dataset. Furthermore, the DTGEN approach outperforms the LSTM-based text generation by 10.11% higher accuracy for the samples in Cornell Movie–dialog Corpus (CMC) dataset.

**Keywords**: *Text Generation, Deep Transfer Learning, Elmo, Context Vector, VAE, Knowledge Transfer, And Bi-LSTM*

## 1. INTRODUCTION

In the real world, Natural Language Processing (NLP) has been emerging as a significant research area due to the rapid increase of the natural language content in various application domains. In NLP, natural language generation is still one of the challenging tasks in producing understandable text from textual data, image data, knowledge bases, and numerical data [1]. Among different natural language generation tasks, text-to-text generation [2] has become one of the potential applications such as dialogue generation, text summarization, and machine translation. Text generation takes the sequence of keywords as the input and generates the output text by processing the input text using different text mining methods [3]. For instance, dialogue generation and question answering generate the interactive responses and answers to the questions, text summarization generates the concise form of the source text with the salient information, and machine translation produces the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

text in a different language for the source text. To precisely develop the text generation solutions, the existing researchers have adopted the deep learning models with the encoder-decoder paradigm [4]. Neural networks have provided promising results in text generation with large datasets. However, generating the absence of the linguistic terms for the sequence of text is an arduous task in the text generation models [5].

Over the past few years, the emergence of transfer learning [6] has gained popular attention in text generation by addressing the shortcomings in the traditional supervised learning models on a single dataset. In transfer learning, the learning model is pre-trained on the large-scale and the generalized text dataset and adapted to the target task of the text generation. Several research works have been developed using the transfer learning models for text generation [7]. However, transfer learning models often confront the transferring of the relevant knowledge from the source domain to the target domain, which leads to the negative transfer. Some research has focused on addressing the negative transfer constraint in the transfer learning model on different complex tasks. Although, improving the generalization ability of the learning model for the target domain with the small amount of training knowledge from the source domain is critical. In addition, the transfer learning-based text-generation models encounter the computation complexity due to the pre-training on the large-scale data source. Even though deep learning models require large-scale data, learning from the significant or relevant training knowledge alone is vital to generate the text for the input textual data [8]. The autoencoder models have been widely utilized in deep learning for the significant representation to reduce the dimensionality of the large-scale data. In the context of text generation, Variational AutoEncoder (VAE) models play a vital role rather than other autoencoder models for the massive input text [9]. Moreover, Recurrent Neural Network (RNN), particularly Long Short-Term Memory (LSTM) deep learning model is suitable for the text classification with the advantage of supporting the text processing for variable length and resolving the vanishing gradient problem in the RNN. It is essential to provide the input feature from the pertinent word embedding model to ensure effective text processing in the LSTM [10]. Thus, this work designs the text generation model with the assistance of the deep transfer learning model. It focuses on avoiding the negative transfer in the transfer learning model through low-dimensional embedding and the VAE-based representation. Moreover, the research work models the text generation with the Bi-LSTM to effectively predict the text sequence.

The major contributions of the proposed Deep Transfer Learning-based Text Generation (DTGEN) work are discussed as follows.

**a.** This research work models the text generation with the deep transfer learning and Bi-LSTM through the contextual embedding and dimensionality reduction for the large-scale text data.

**b**.Initially, the proposed DTGEN approach applies the Embedding from Language Model (ELMo) to generate the contextual vector representation of the input text sequence.

**c**.In the design of the deep transfer learning-based text generation model, it generates the low-dimensional vector representation with the assistance of the weighted regularization in the VAE model that aids to filter the relevant training knowledge by mapping the contextual vector from the source domain.

**d**.Moreover, the proposed DTGEN approach designs the prediction model with the Bi-LSTM for learning the sequential patterns of the input text from both directions to generate the relevant terms for the text sequence.

**e**.Thus, the experimental results illustrate that the proposed DTGEN approach significantly outperforms the existing text generation approach with the advantage of deep transfer learning, contextual ELMo-based representation, VAE-based dimensionality reduction, and Bi-LSTM based text prediction.

### 1.1. Paper Organization

The structure of this paper is organized as follows. Section 2 reviews the related works in the feature extraction for the text generation and the neural network-based text generation model. The overview of the research contributions and processes involved in the proposed DTGEN methodology are presented in Section 3. Section 4 describes the proposed text generation methodology DTGEN in detail. The evaluation model and result analysis for the text generation model are presented in Section 5. Section 6 discusses this research work and its results with a detailed analysis. Finally, Section 7 concludes this work, and the outlook of this work is described in Section 7.1.

## 2. RELATED WORKS

This section reviews several previous text generation research works, including the different feature engineering and deep learning-based approaches.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

### 2.1. Feature Engineering in Text Generation Approaches

Texygen model [11] standardizes the text generation through the fine-tuned implementation to evaluate the quality, consistency, and diversity of the generated texts. The data-to-text generation model [12] incorporates the content selection and planning stages as the generation of content plan with the potential information and the document generation based on the content plan, respectively. By presenting the neural network architecture, it resolves the shortcomings in the accurate content selection without redundant text generation from the analysis of the input data and conditions of the content plan. However, it fails to cope with the knowledge across different domains while generating the content of the text. In consequence, it shows that there is an essential need to utilize the domain knowledge to generate the text sequence with high quality. The controllable content selection approach [13] presents the task-agnostic model and the end-to-end training with a lower bound of marginal likelihood maximization. It decouples the content selection from the decoder and maintains the trade-off between the controllability and performance in the text generation. Even though it ensures the text with diversity and controllable content selection, the objective of the text generation varies with respect to the applications, and hence, training the model with the massive textual content leads to inaccurate content selection for the text generation. From the study of the controllable content selection work [13], it is recognized that the need for the low-dimensional representation for the text generation. The research work [14] presents a MoverScore metric to validate the text generation tasks such as machine translation, summarization, data-to-text generation, and image captioning. It combines the semantic distance measure with the contextualized representations to improve the quality of evaluating the text predictions. Combining the contextual embedding with the Earth Mover's distance facilitates the measure of the text generation quality; however, the semantic deviation on the different domain knowledge reduces the generalization capability. Hence, it induces the development of the contextual knowledge transfer from the relevant source domain for the text generation in the proposed research procedure. The natural language generation model [15] adapts transfer learning as a large-scale pre-trained language model to analyze the different types of input context with different inductive biases. By modeling the encoder-decoder structure, it processes the input contexts and generates the text sequence as the output. However, it lacks to focus on contextually filtering and transferring the knowledge from the source domain to the target domain. As a result, it paves the way to develop the contextual knowledge transfer in the transfer learning model for the quality of the text generation. Discourse Representation Structure (DRS)-based text generation model [16] addresses the variable naming and condition ordering during the text generation using the sibling tree LSTM model. It encodes the rhetorical relations of the semantic information, co-reference, and presupposition within and across sentences in a document-level for the text generation. Although, lack of contextual embedding leads to the building of the sentence with the discourse representation regardless of the context of the individual term with the associated terms in the sentence. Hence, the proposed approach intends to generate the low-dimensional representation of natural language text for sequential pattern learning.

### 2.2. Deep Learning-based Text Generation Approaches

Semantically Controlled-LSTM (SC-LSTM) model [17] generates the natural language by jointly optimizing the components of surface realization and sentence planning. By modeling the cross-entropy training criterion, it generates linguistically varied responses even when learning from the unaligned data. Instead of contextual word embeddings, pre-trained word vectors misleads the representation or vectorization of the non-grammatic natural language texts due to the diversified sentence representation.Consequently, the proposed approach targets designing the text generation model with the contextual vector representation. Affect-LM model [18] generates the customized conversational text based on the degree of emotional content by enriching the Long Short-Term Memory language model. By learning the affect-discriminative word representations, it improves the text prediction without compromising the grammatical structure of the sentences. However, the scarcity of domain knowledge regarding the textual content fails to generate domain-relevant keywords and the terms related to the emotions expressed in the text. Topic-Guidance-based text generation approach [19] models the Topic-Guided Variational AutoEncoder (TGVAE) with the Gaussian Mixture Model (GMM) to generate the sentences based on the topic. By jointly learning the neural topic module and conditional VAE-based neural sequence module, it generates the text from the clustering structure regardless of the predefined labels. The unsupervised learning model along with the topic knowledge is inadequate to generate the content for the natural language text. From the analysis of the existing research works [18, 19], the proposed approach targets adopting the deep transfer learning model with the contextual knowledge transfer to address the data scarcity. The structural neural encoder model [20] generates the Abstract Meaning Representation (AMR) graph with the long-range dependencies and reentrancies. By combining the BiLSTM network and GCNwired, it improves the overall text generation performance than the sequential and tree encoders. Even though bi-directional learning-based meaning representation generates the sequential text, it confronts the concise text generation from the massive collection of textual content.In consequence, it is determined that feature extraction and representation become a vital task in the text generation process. Knowledge-Grounded Pre-Training (KGPT) model [21] initially generates the knowledge-enriched text and fine-tuned task-specific text. It effectively generates text under few-shot and zero-shot settings by utilizing the knowledge from the unlabeled external data. However, learning the text data in the low-dimensional space is critical for the knowledge-enriched text, leading to a massive amount of enriched terms in the generated content. The plug and Play approach [22] mitigates the need for the labeled data for the text generation through the embedding-to-embedding within the autoencoder. By learning the text in the low-dimensional embedding space, it ensures effective text generation from the benefit of the adversarial loss term and the neural mapping architecture. Moreover, the plug and play model directly implements the attribute transfer and content preservation on the text autoencoder based on two conditional generation models for the supervised style transfer and unsupervised style transfer. Although, low-dimensional embedding space with the contextual embedding generates the potentially key terms in the generated content, which impacts the generation of redundant terms in different contexts. From the analysis of shortcomings in the research works [21, 22], it is crucial to represent the low-dimensional vector spaces with the contextual features to generate the non-redundant text sequence. Context-based text generation model [23] trains the LSTM language model with the contextual information extracted from the semantic meaning of the sentences in the word vector spaces. By examining the semantic relations among the input text, context vector, and target text, it generates the text based on the word clustering and word importance variations. However, it requires external knowledge to generate the contextual sentence with the sequential relations for the natural language text. The semi-supervised text generation approach [24] addresses the data scarcity constraint through the pre-trained encoder-based data augmentation and the transfer learning by training the deep learning-based code-mixed text generator. It augments the encoder in the code-mixed text generator with the task-agnostic and linguistic features and transfers the knowledge from the Neural Machine Translation (NMT).However, the lack of considering the negative transfer constraint in the transfer learning model leads to inaccurate textual content generation. Thus, to improve the quality of the text generation, it is significant to contextually transfer the source domain knowledge to the target domain without compromising the training time. The deep successor feature learning model [25] generates the sequences using reinforcement learning along with the value functions of the reward predictor and successor map. By combining the encoder-decoder model with the successor feature model, it learns the semantic meaning of the utterances in the sentences with the optimization of the long-term goals.Even though the learning model with the attention mechanism leverages the text generation, the lack of contextual feature representation in the patterns for the learning model misleads the sequence generation with relevant terms. The next word prediction approach [26] applies the Bi-LSTM model to predict the sequential words in the Hindi language. By suggesting the next few words for the sequence, it reduces the number of keystrokes entered by the user and mitigates the spelling mistakes in the generated sentence. However, it fails to generate the new and relevant terms that do not exist in the vocabulary due to the lack of learning the new vocabularies from the external knowledge source while suggesting the personalized words. As a result, the proposed work focuses on modeling the text generation with the contextual embeddings for low-dimensional vector representation and contextual knowledge transfer in deep transfer learning.

Most of the natural language generation tasks have been widely applied in the sequence encoder-decoder architecture to generate the sequence from analyzing intrinsic features in the sequence-to-sequence models. Although, capturing the context of the sequence is challenging to the text generation model due to the mapping of the word sequences alone to the successive word. Lack of adequate contextual information in the input text data leads to the generation of the sentence without the pragmatic sense even generating the grammatically stable sentence. In addition, text generation models confront the inadequate domain knowledge to generate the semantically associated terms and new terms. Lack of contextual embedding leads to the inaccurate generation of the text sequence for the non-grammatical representation of the natural language text. Even though transfer learning methods have addressed the data scarcity for the text generation, the lack of transferring or utilizing the relevant knowledge alone from the source domain misleads the inaccurate text generation. Learning the input embedding patterns from the high-dimensional data space misguides the understanding of the contextual terms in the sentence due to the increased number of ambiguated terms in the natural language text. Hence, it is essential to design the deep transfer learning-based text generation model with the addressing of negative knowledge transfer and consideration of the contextual embedding.

## 3. An Overview Of The Proposed Model

Natural language processing has become one of the significant research areas in the field of text mining. Natural language generation plays a prominent role in the real world owing to the rapid access of the Web through unstructured text. Text generation, auto-captioning of images, and summarization are the natural language understanding applications. The main goal of automatic text generation is to generate new text from analyzing the sequence of previous words given by the users using deep learning techniques. Despite this, deep learning-based text generation researches confront the understanding of the context of the sentence due to the accumulation of the text content with most of the irrelevant or non-significant words. Hence, several researchers have focused on applying the pre-trained models to recognize the patterns from the input sentence accurately. Moreover, due to the need for a large amount of data and wide coverage of knowledge for the analysis of the textual content, several research works have applied the transfer learning methods that utilize the information from the source domain regarding the context of the sentence and apply the learned data in the target domain for the text generation. In essence, to effectively extract the knowledge from the source domain, there is an essential need for determining the relevant context discussed in the textual content, which is accomplished by the dimensionality reduction method. **Figure 1(Pg.No.21)** illustrates the steps involved in the proposed text generation methodology.

**Step1:Text Preprocessing and Text Vectorization:** The proposed DTGEN approach conducts the preprocessing on the raw natural language text to grasp the potential insights for the task of precise text generation. During the preprocessing, it applies the lemmatization, lower or upper case conversion, number conversion, abbreviation expansion, and special characters removal. To contextually provide the vector representation for the input text sequence, the proposed DTGEN approach models the text vectorization with the assistance of the ELMo. In essence, it represents the sentences with the accumulation of the contextual word vectors.

**Step2:Low-Dimensional Text Feature Modeling:** The proposed DTGEN approach adopts the VAE model to represent the low-dimensional text data to enhance the performance of the deep transfer learning model. The increased dimensionality of the input sentence increases the structure of the text, which leads to the inaccurate mapping of the input data with the source information during the learning by the deep learning model. Hence, transforming the high-dimensional text data into low-dimensional vectors is essential for the automatic text generation system.

**Step2.1.Contextual Text Sequence Representation:** The proposed DTGEN approach utilizes the context vector from the embedding model to extract the low-dimensional data from the source domain in the deep transfer learning model. Applying the deep transfer learning with dimensionality reduction ensures the text generation even when there is a lack of knowledge on the domain information and irrelevant textual words in the input sentence. Hence, the proposed approach assists in transferring the contextual knowledge alone from the massive source text by mapping the context vector with the text sequences in the source domain. The DTGEN approach updates the weighted regularization in the VAE model during the representation of the text sequence to reduce the dimensionality of the data. As a result, the DTGEN approach provides the contextual and low-dimensional feature as the input to the deep transfer learning model.

**Step3:Deep Transfer Learning-based Text Generation:** To effectively generate the text sequence for the target domain, the proposed DTGEN approach utilizes the contextual sequence representation as to the pre-trained knowledge in the deep transfer learning model.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

**Step3.1:Sequence Analysis through Transfer Learning:** By mapping the context vector, source domain knowledge, and target domain knowledge, the proposed DTGEN approach conflates the knowledge for pre-training the transfer learning model with the ELMo. Consequently, it ensures the knowledge transfer contextually based on the word-level and sentence-level contextual embeddings. Instead of learning the represented feature independently, the proposed DTGEN approach examines the sequential representation of the context vectors in the source domain to predict the relevant terms or sequence for the target domain.

**Step3.2:Bi-directional Sequence Learning:** The proposed DTGEN approach predicts the text sequence with the contextual analysis of the input sequence. In addition to the contextual and low-dimensional representation of the text sequence, it models the text generation with the Bi-LSTM model. The Bi-LSTM learns the relationships among the text patterns from both the forward and backward directions to predict the new word in the sentence effectively. Hence, to predict the sequence of new and relevant terms, the proposed DTGEN approach applies the Bi-LSTM network to associate the ELMo-based embedding representation and contextually transferred knowledge in deep transfer learning.

**Step4:Text Sequence Prediction:** Finally, the proposed DTGEN approach generates the text sequence through deep transfer learning and Bi-LSTM based prediction with the advantage of the inherent sequential relationship analysis in forward and backward directions.

## 4. The Proposed Dtgen Methodology

With the target of developing the text generation model for the natural language text even when there is a lack of adequate knowledge in the input text, this work enriches the procedures involved in the existing deep learning-based text generation process. It essentially enhances the vectorization of the input text without the irrelevant information through ELMo based low-dimensional vector representation. Moreover, the research work enriches the deep transfer learning with the contextual knowledge transfer by utilizing the low-dimensional representation of the text instead of learning the pattern from the vectorized output of the massive input textual sequence. Also, it averts the transferring of irrelevant and abundant knowledge to the target domain during the generation of text sequence. As a result, the research work enriches the text generation task with the assistance of a deep learning model. **Figure 2 (Pg.No.2770)**shows the overall process involved in the proposed methodology.

### 4.1. Sequence Representation through Dimensionality Reduction

With the target of representing the sentence with contextual features, the research work models feature engineering with the assistance of the contextualized embedding and unsupervised deep learning model. The proposed methodology models text vectorization with the assistance of Embedding from Language Model (ELMo) and sequence representation using dimensionality reduction.

#### 4.1.1. Vectorizing the Text Data

Initially, the DTGEN approach applies text preprocessing on the text data before performing the text featurization to generate meaningful data from the raw data containing numerical values, special characters, and abbreviations. The DTGEN approach transforms the unstructured text data into the structured form of a word embedding vector to extract the subjective information contained in the input text. After performing the text normalization, the DTGEN approach models the ELMo to convert the large text data into vector space. By determining the meaning of the words in their location in the feature space, the ELMo vectorizes the words of the vast amount of data. With the rapid increase of the text dimensions in the natural language text, encoding the words as the vectors with the knowledge of the context of the word ensures the effective understanding of the text. In essence, the DTGEN approach vectorizes the sentence instead of vectorizing a single word in the sentence with the assistance of the ELMo neural network model.

$$\mathrm{W_i} = \alpha^{k}\left[\sum_{S}\sum_{j=0}^{L}\left(\mathrm{SW_j^k} \times \mathrm{h_{i,j}^{LM}}\right)^{S}\right] \quad (1)$$

The DTGEN approach computes the task-specific weighting for all the bidirectional language models using equation (1), which enforces the generation of representation concerning the context of the word. ELMo word representation is the function of the input sentence, can learn the syntactic and semantic features and the dynamicity in the lexical usages. By aggregating all the information of the words in a sentence as a single vector, the DTGEN approach recognizes the context-dependent information as well as the syntax of the tokens in

Journal of Theoretical and Applied Information Technology
15th May 2022. Vol.100. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the sentence (S). In equation (1), $\alpha^k$ and $SW_j^k$ represents the weight of the word vector and softmax-normalized weight in the $j^{th}$ layer for the task (k), respectively. In addition to the contextual representation of the word vector, the proposed approach considers the sentence with the multiple context word vectors to generate the representation for each sentence, avoiding the complexity in addressing the discourse relations in the text sequences.

$$P(w_1, w_2, \ldots, w_N) = \prod_{i=1}^{N} P(w_i | w_1, w_2, \ldots, w_{i-1}) \quad (2)$$

$$P(w_1, w_2, \ldots, w_N) = \prod_{i=1}^{N} P(w_i | w_{i+1}, w_{i+2}, \ldots, w_N) \quad (3)$$

By utilizing a bidirectional language model of the Bi-LSTM, the ELMo-based DTGEN approach generates the word representation for the sequence of 'N' words. In the DTGEN approach, the Bi-LSTM model predicts the probability of each word ($w_i$) in the context based on the probability of the sequence of the words in the sentence through forward and backward analysis. Equations (2) and (3) compute the probability of the word ($w_i$) according to the context (k) in the forward layer of the language model and the backward layer of the language model, respectively. Thus, the DTGEN approach converts each word with its surrounding words in a sentence with the assistance of the ELMo into low-dimensional vectors.

#### 4.1.2. Representing the Vectorized Sequence

The DTGEN approach examines the vectorized input sentence to structure the feature vectors in a low-dimensional space. Instead of representing the sentence and reducing the dimensions through word embedding, the DTGEN approach transforms the high-dimensional vector space into low-dimensional vector space using the autoencoder without sacrificing the performance of the algorithm. The dimensionality reduction plays a significant role in extracting the potential information from the source domain in transfer learning. By removing the noisy, redundant, and irrelevant features in the input data, the DTGEN approach enhances the performance of the underlying transfer learning model. Even though contextual word embedding model becomes the fundamental building process in text mining natural language processing, further improving the representation of the pre-trained word vectors is essential. The word embeddings with dimensionality reduction facilitate the reduction in the size of word embeddings for the sentences, paragraphs, or documents. Hence, the DTGEN approach enhances the quality of the word embedding with more discriminative using the variational autoencoder due to the generation of different vector representations for the same word regarding the context of the word in the sentence. Variational autoencoder is the regularized version of the autoencoders to model the generative process. In essence, the DTGEN approach models the autoencoder as the post-processing algorithm to reduce the dimension of the pre-trained word embeddings. Equation (4) computes the loss in the variational autoencoder based on the reconstruction quality and the weighted regularization term.

$$L(\theta, \emptyset; x_i) = E_{Q_\emptyset(z^{LT}|x_i)}[\log P_\theta(x_i|z^{LT})] - \frac{W}{LT} * D_{KL}\left(Q_\emptyset(z^{LT}|x_i) \| P(z^{LT})\right) \quad (4)$$

In equation (4), θ and φ refer to the decoder and encoder parameters, respectively. According to the variational autoencoder model, the DTGEN approach computes the loss function to adopt text data regeneration or text modeling. 'LT' indicates the length of the text or timesteps, and 'W' denotes the weight for the KL regularization term. By examining the embeddings of a subset of words at a time, the DTGEN approach generates the low-dimensional sequential representations in an online fashion along with regularization and non-linear transformations. Thus, the DTGEN approach enriches the isotropy of the pre-trained word embeddings by applying the variational autoencoder.

### 4.2. Generating Text with Deep Transfer Learning

The DTGEN approach generates the text from the learning of low-dimensional word embeddings using deep transfer learning with Bi-LSTM. The DTGEN approach adaptively transfers the relevant knowledge in terms of text terms from the source domain to the target domain through parameter tuning to resolve the negative transfer constraint in the transfer learning model. By contextually transferring the knowledge, it ensures the accurate prediction of the texts for the input data sequence.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

**4.2.1. Contextual Knowledge Transfer and Sequence Generation**

The DTGEN approach models the text generation with the contextual knowledge transfer to deal with the lack of knowledge on the generation of the texts. By integrating the knowledge from pre-trained word embeddings with the unlabeled data, it enhances the deep transfer learning-based text generation. The DTGEN approach ignores the irrelevant source domains while transferring the knowledge for the target domain through adaptive learning rate updation. With the measure of the similarities on the feature distribution of the source and target domains, it determines that the potential keywords or features are to be transferred to the target domain to facilitate the text generation. Initially, the DTGEN approach trains the deep transfer learning model on the task-specific or domain-relevant text data and fine-tunes the language model for the input text sequences in the target domain to generate the text with the enriched set of vocabularies. The contextual knowledge transfer in the DTGEN approach extracts the context vectors to bridge the sequence preservation between the input and output text sequences.

$$p(\mathrm{ST_u}|\mathrm{C_i}) = \begin{cases} 0.5\,, \text{if ST} = \text{TT and ST} \neq \text{C} \\ 1, \text{if ST} = \text{TT and ST} = \text{C} \\ 0, \text{if ST} \neq \text{TT} \end{cases} \quad (5)$$

By applying equation (5), the DTGEN approach retains the semantic information of the sequence based on the context vector during the knowledge transfer. In equation (5), ST, C, and TT refer to the source text, the context of the input sentence, and target text, respectively. $ST_u$ and $C_i$ denotes $u^{th}$ source domain and $i^{th}$ context sentence or vector. The proposed approach assigns the significance score as 0, 0.5, and 1 in transferring the knowledge from the source domain to the target domain based on the mapping of the knowledge. If source text and target text are in different contexts or vary in their domains, the DTGEN approach ignores the transferring of the knowledge from that source domain. Whereas, if the source text and target text are matched in the same domain along with the context matching from the input sentence, it provides higher priority or weight to the knowledge transfer; otherwise, it partially transfers the knowledge from the source domain when there is context not matched with the source domain. In the DTGEN approach, the generation of text for every sequence of inputs heavily relies on the context vector-based training in the deep transfer learning model. The DTGEN approach examines the impact of the hyperparameter, especially the learning rate, in the accuracy of the learning process to train the transfer learning model over the large-scale data efficiently. By optimally tuning the learning rate, the DTGEN approach enforces the precise text generation in the target domain. If the learning rate is too high, the learning model requires multiple iterations to reach the convergence level. At the same time, a low learning rate leads to the local convergence instead of finding the optimal point globally. Thus, the DTGEN approach fine-tunes the deep transfer learning model with the optimal learning rate during the analysis of the text data in the target domain with the reference of the filtered information in the source domain.

Moreover, the DTGEN approach models the Bi-LSTM model to learn the input sequence from forward and backward directions based on equations (2) and (3). The DTGEN approach predicts the contextual target y= $\{y_1, y_2,...,y_t\}$ for the input sequence of input x=$\{x_1, x_2,...,x_s\}$. By applying the conditional distribution for the output contextual target ($y^c_{target}$), it utilizes the input data ($x_c$) including the source input (x) in the form of a context vector along with the contextual knowledge (C) transferred from the source domain.

$$p\left(y^{c}_{target}\middle|x\right) = \prod_{w_i=1}^{N_c} p\left(y_{w_i}\middle|y_{w_{i-1}}, x_c\right) \quad (6)$$

In Equation (6), $w_i$ and $N_c$ refer to the keywords or terms and keywords in the contextual keywords transferred by the deep transfer learning model. The DTGEN approach augments the text generation with the knowledge of the contextual transfer. In essence, it analyzes the contextual vector form of the input sequence in both the forward and backward directions and utilizes the additional contextual knowledge from the source domain to generate the text for the target domain with the assistance of the deep transfer learning model. Consequently, the DTGEN approach dynamically updates the knowledge contextually and generates the relevant terms in the text sequence without predicting the texts from the terms associated with the target domain alone. Thus, it effectively generates the text for the different input sequences referring to different contexts with the adaptive transfer learning and Bi-LSTM model.

Journal of Theoretical and Applied Information Technology
15th May 2022. Vol.100. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

## 5. Experimental Evaluation

To exemplify the performance of the text generation model, the experimental model evaluates the proposed DTGEN approach with the comparison of the existing Deep Learning-based Next Word Prediction (DNWP) approach [26] and the baseline algorithms as LSTM and Bi-LSTM models while testing on benchmark text dataset.

### 5.1. Experimental Setup

During the implementation of the text generation model, the experiments are to be conducted on Ubuntu 16.04 64-bit machine with a 3GHz Intel CPU and 16GB memory. The experimental model implements the text generation algorithm using the Python programming language. It runs the python libraries with the python version 3.6.8 to implement the deep learning model, including the pandas, numpy, sklearn preprocessing, text processing libraries, and sklearn metrics. To test the text generation model, the experimental framework utilizes two different text datasets, such as the Newsroom dataset [27] and Cornell Movie–dialog Corpus (CMC) dataset [28]. The newsroom dataset consists of 1.3 million articles and summaries, widely used to train and evaluate the summarization system. The articles and summaries are created by the editors and authors in the newsrooms of 38 major publications. In the CMC dataset, the movie corpus comprises a set of fictional conversations gathered from the movie scripts. The CMC dataset contains 220,579 conversational texts between 10,292 pairs of movie characters.

#### 5.1.1. Evaluation Metrics

The experimental model utilizes the following evaluation metrics to assess the performance of the text generation model.

**Bilingual Evaluation Understudy (BLEU):** BLEU is a precision-focused metric that computes the n-gram overlap between the generated text and the reference text. BLEU score relies on the evaluation of the n-gram overlap.

**Recall Oriented Understudy for Gisting Evaluation (Rouge):** Rouge is a recall-focused metric that computes the n-gram overlap between the generated text and the reference text through n-rouge, l-ROUGE, and s-ROUGE.

**Accuracy:** It is the percentage of overall correctly predicted n-grams in the generated text over the n-grams in the reference text.

**Loss:** It denotes the categorical cross entropy-based loss of the text generation model, based on the mapping between the actual n-gram or word to be predicted and the predicted n-gram or word.

### 5.2. Experimental Results

This section illustrates the evaluation results of the proposed DTGEN, existing DNWP, Bi-LSTM, and LSTM models while testing on the Newsroom and CMC datasets for the different epochs.

#### 5.2.1. Number of Epochs Vs. BLEU

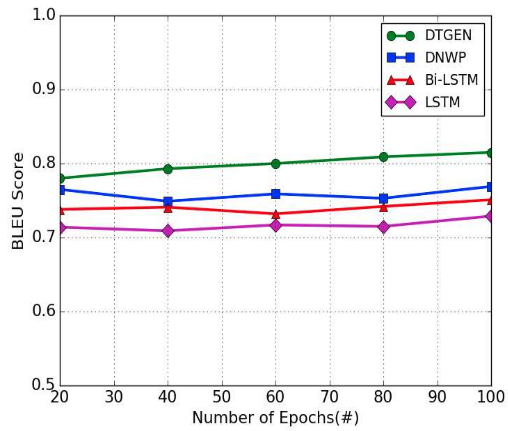

*Figure 3: Comparison of BLEU Score*

**Figure 3** shows the comparative performance of the BLEU score for the proposed DTGEN, existing DNWP, and baseline algorithms of Bi-LSTM and LSTM models while testing on the Newsroom dataset. The proposed DTGEN approach linearly increases the BLEU score from 0.78 to 0.815 while varying the number of epochs as 20, 40, 60, 80, and 100. Compared to the existing DNWP approach and the Bi-LSTM based text generation model, the DTGEN approach accomplishes 0.046 and 0.064 higher BLEU scores for 100 epochs. It is because the proposed DTGEN approach generates the contextual vector and context vector-based knowledge extraction using the VAE model in addition to the Bi-LSTM based text prediction process. As a result, the DNWP approach obtains a minimal BLEU score with the difference of 0.056 BLEU score for 80 epochs even the Bi-LSTM model assists in accurately predicting the next words in the Newsroom dataset.

#### 5.2.2. Number of Epochs Vs. ROUGE

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

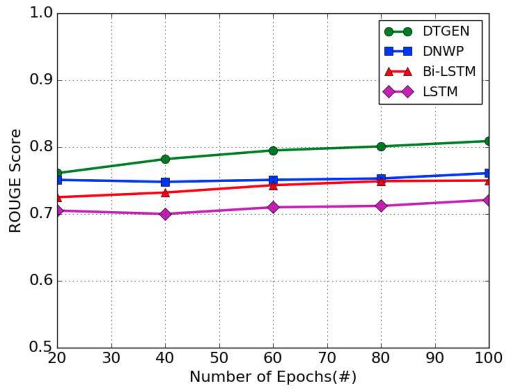


*Figure 4: Comparison of ROUGE Score*

Figure 4 illustrates the ROUGE score for the proposed and existing text generation approaches, along with the comparison of baseline algorithms of Bi-LSTM and LSTM models for the CMC dataset. When epochs are 100, the proposed DTGEN approach accomplishes 0.048 and 0.059 higher ROUGE scores than the existing DNWP and Bi-LSTM models. The DTGEN approach models the deep transfer learning with the contextual knowledge transfer and applies the weighted regularization-based VAE model ensures adequate knowledge pre-trained in the learning model. As a result, the DTGEN approach avoids generating Out-Of-Vocabulary (OOV) words and redundant words due to the training on abundant relevant terms in the source domain. Even though the Bi-LSTM based prediction model in the DNWP approach, creating the dictionary with unique words misguides the learning model without contextual relations in the text sequence. Hence, the DNWP approach obtains 4.4% performance degradation than the proposed DTGEN approach model when epochs are 60.

**5.2.3. Number of Epochs Vs. Accuracy**

The comparative performance of the accuracy for the proposed DTGEN, existing DNWP, and deep learning algorithms of Bi-LSTM and LSTM models while testing on a single dataset and the different datasets are depicted in Figure 5.

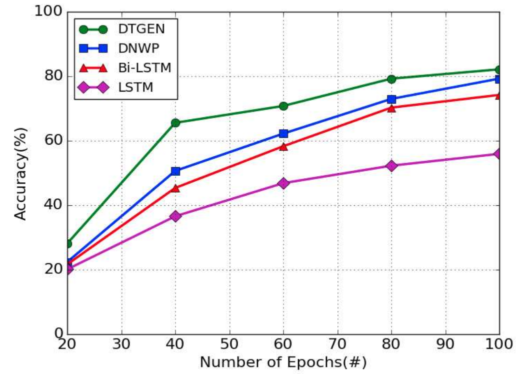


*Figure 5(a): Accuracy of Different Models in Newsroom*

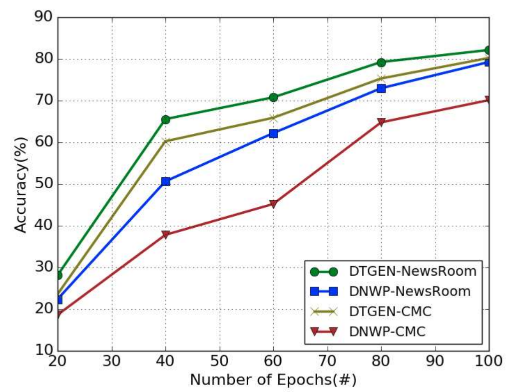


*Figure 5(b) Accuracy on Newsroom and CMC*

Figure 5(a) illustrates the accuracy performance of the proposed DTGEN approach with the comparison of the existing DNWP, Bi-LSTM, and LSTM models for the different number of epochs while testing on the Newsroom dataset. When the number of epochs is 80, the proposed DTGEN approach accomplishes the maximal accuracy of 79.25% with the difference of 2.9% only. In the same scenario, the DNWP approach achieves 72.97% with the difference of 6.29% to reach its better accuracy in 100 epochs. The proposed DTGEN approach consistently predicts the text sequence globally in all iterations by utilizing the context vector and modeling the contextual knowledge transfer with the assistance of the VAE model. Figure 5(b) depicts the comparative performance of the accuracy of the proposed DTGEN, and existing DNWP approaches testing on the text datasets of the Newsroom and CMC. With the massive amount of text data accumulation in the Newsroom dataset, both the DTGEN and DNWP approach accomplishes higher accuracy in the text generation compared to the CMC dataset. The CMC dataset meets the data scarcity constraint, which is addressed by the deep transfer learning in the DTGEN approach. As a result, the DTGEN approach in Newsroom and CMC dataset yields 12.03% and 10.11% higher accuracy than the DNWP approach in the CMC dataset for 100 epochs.

### 5.2.4. Number of Epochs Vs. Loss

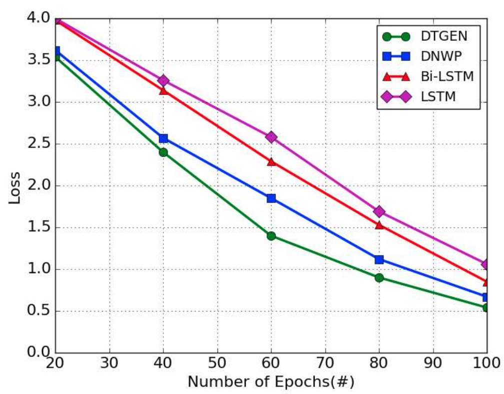


*Figure 6: Comparison of Loss*

Figure 6 shows the loss of the DTGEN, DNWP, Bi-LSTM, and LSTM approaches with the variation of the different epochs from 20 to 100 with an interval of 20 for the Newsroom dataset. For 100 epochs, the DTGEN approach obtains loss value as 0.54 only but, the Bi-LSTM and LSTM models based text generation yields the loss value as 0.85 and 1.06. Compared to the LSTM model, the Bi-LSTM model effectively learns the sequential patterns of the sentences from the forward and the backward directions, ensuring the prediction of the relevant terms. In contrast, the proposed DTGEN approach contextually learns the sequence patterns with the help of the low-dimensional feature representation and the advantage of deep transfer learning. Hence, the proposed approach comparatively obtains a minimal error value even when the Bi-LSTM model is significant in the text generation process.

## 6. Discussion

With the rise of virtual assistants in recent years, accomplishing task completion in various real-time applications through natural language interfaces on the Web has become essential. Due to the emerging need of responding the people all over the world through assistants, the research on text generation has gained significant attention among text mining researchers. This work has been focusing on improving the accuracy of the text generation task with limited domain knowledge. Hence, the proposed work has integrated the contextual word embedding model and deep transfer learning model.

This work showed that our proposed text generation model successfully generates the text sequence for the natural language text through the deep transfer learning model. It was found that the sequence of the text generation heavily relies on the contextual information of the available text, analyzed from the research works [15, 16]. The text generation is highly sensitive to the sharp changes in the contextual features in the text. From the result of low-dimensional contextual feature representation and deep transfer learning-based sequence generation, our model predicts the text sequence for the context of the sentence. The performance of the reduced representation of the contextual embeddings matches with the text generation from the high-dimensional representation [21], showing the significance of feature dimensionality reduction. Because contextual embedding is a vital task in the understanding of the natural language, resulting in gains in accuracy through the low-dimensional representation, it proves useful to the applications that interact with the natural language interface. This work obtains comparatively better performance on the text generation because the proposed algorithm effectively applies the deep transfer learning with the contextual ELMos embeddings to potentially provide the relevant knowledge alone in the sequence representations and learn the contextual patterns for sequence prediction. Although, the reduced embeddings suffer from negligible performance degradation in terms of accuracy in the CMC dataset due to the scarcity constraint. In addition, contextual knowledge without adaptively modeling the network structure instead of the learning rate updation leads to a decline in the performance of accuracy on different text generation datasets.

## 7. Conclusion

This paper presented the deep transfer learning-based text generation model, DTGEN model. The DTGEN approach has applied the ELMo embedding model, VAE model, and Bi-LSTM model to generate the text considering the contextual representation of the textual terms. To model low-dimensional feature representation of the input text sequence, the proposed approach has adopted the ELMo and transformed the raw input data into the contextual vector. In subsequence, to avoid the negative transfer constraint in the transfer learning, it has designed the VAE with the weighted regularization during the pre-training of the learning model on the source domain with the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

reference of the contextual vector. Moreover, the DTGEN approach has predicted the textual terms for the sequence in the target domain using the Bi-LSTM model. Thus, the experimental results illustrate that the DTGEN approach significantly outperforms the existing DNWP approach in terms of accuracy by 2.89% while testing on the Newsroom dataset. In the future, this work will extend to the modeling of the text generation with the discourse representation. Furthermore, attention-based deep learning models can be adopted for the natural language text generation in the future.

### 7.1. Outlook

Over the past decades, the natural language generation task has gained significant attention and advancements in research. This paper has sought to resolve the shortcomings in text generation even when there is a lack of adequate knowledge in the natural language text. Initially, we have studied the recent developments in text generation to analyze the research gaps in natural language generation applications extensively. From the analysis of the emerging related text generation solutions from the traditional research, the text generation task substantially impacts the domain-dependent and domain-independent knowledge instead of depending on the available text data. In the conventional text generation models, maintaining a trade-off between the efficiency and quality of the generated text becomes a research constraint. Compared to the rule-based text generation approaches, data-driven approaches are efficient; however, the quality of the text generation is likely to be compromised. At the conclusion of this research, it is determined that there is a fast pace of the developments required for both the academia and industry fields with the potential research directions. The research was performed towards enhancing the text generation process with a fundamental interest in the natural language generation and the technological importance in the real world.

## REFERENCES:

Khurana, Diksha, Aditya Koli, Kiran Khatter, and Sukhdev Singh, "Natural language processing: State of the art, current trends and challenges", *arXiv preprint arXiv:1708.05148, 2017.*

[2] Yang, Diyi, and Lucie Flek, "Towards User-Centric Text-to-Text Generation: A Survey", In *International Conference on Text, Speech, and Dialogue*, pp.3-22, Springer, Cham, 2021.

[3] Yu, Wenhao, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang, "A survey of knowledge-enhanced text generation", *arXiv preprint arXiv:2010.04389, 2020.*

[4] Iqbal, Touseef, and Shaima Qureshi, "The survey: Text generation models in deep learning", *Journal of King Saud University-Computer and Information Sciences, 2020.*

[5] Jin, HanQi, Yue Cao, TianMing Wang, XinYu Xing, and XiaoJun Wan, "Recent advances of neural text generation: Core tasks, datasets, models and challenges", *Science China Technological Sciences*, pp.1-21, 2020.

[6] Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He, "A comprehensive survey on transfer learning", *Proceedings of the IEEE*, Vol.109, No.1, pp.43-76, 2020.

[7] Ruder, Sebastian, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf, "Transfer learning in natural language processing", In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:* Tutorials, pp.15-18, 2019.

[8] Tan, Chuanqi, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu, "A survey on deep transfer learning", In *International conference on artificial neural networks*, pp.270-279, Springer, Cham, 2018.

[9] Zhai, Junhai, Sufang Zhang, Junfen Chen, and Qiang He, "Autoencoder and its various variants", In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp.415-419, 2018

[10] Wang, Shirui, Wenan Zhou, and Chao Jiang, "A survey of word embeddings based on deep learning", *Computing,* Vol.102, No.3, pp.717-740, 2020.

[11] Zhu, Yaoming, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu, "Texygen: A benchmarking platform for text generation models", In The *41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp.1097-1100, 2018.

[12] Puduppully, Ratish, Li Dong, and Mirella Lapata, "Data-to-text generation with content selection and planning", In Proceedings of the *AAAI conference on artificial intelligence*, Vol.33, No.01, pp.6908-6915, 2019

[13] Shen, Xiaoyu, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine, "Select and attend: Towards controllable content selection in text generation", *arXiv preprint arXiv:1909.04453, 2019*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[14] Zhao, Wei, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger, "Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance*", arXiv preprint arXiv:1909.02622, 2019*

[15] Golovanov, Sergey, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf, "Large-scale transfer learning for natural language generation", In Proceedings of the *57th Annual Meeting of the Association for Computational Linguistics*, pp.6053-6058, 2019

[16] Liu, Jiangming, Shay B. Cohen, and Mirella Lapata, "Text Generation from Discourse Representation Structures", In Proceedings of the *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pp.397-415, 2021.

[17] Wen, Tsung-Hsien, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems", *arXiv preprint arXiv:1508.01745, 2015.*

[18] Ghosh, Sayan, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer, "Affect-lm: A neural language model for customizable affective text generation", arXiv preprint *arXiv:1704.06851, 2017.*

[19] Wang, Wenlin, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin, "Topic-guided variational autoencoders for text generation", *arXiv preprint arXiv:1903.07137, 2019.*

[20] Damonte, Marco, and Shay B. Cohen, "Structural neural encoders for AMR-to-text generation", *arXiv preprint arXiv:1903.11410, 2019.*

[21] Chen, Wenhu, Yu Su, Xifeng Yan, and William Yang Wang, "KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation", *arXiv preprint arXiv:2010.02307, 2020.*

[22] Mai, Florian, Nikolaos Pappas, Ivan Montero, Noah A. Smith, and James Henderson, "Plug and Play Autoencoders for Conditional Text Generation", *arXiv preprint arXiv:2010.02983, 2020.*

[23] Santhanam, Sivasurya, "Context based text-generation using lstm networks", *arXiv preprint arXiv:2005.00048, 2020.*

[24] Gupta, Deepak, Asif Ekbal, and Pushpak Bhattacharyya, "A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning", In Proceedings of the *2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp.2267-2280, 2020

[25] Xu, Cong, Qing Li, Dezheng Zhang, Yonghong Xie, and Xisheng Li, "Deep successor feature learning for text generation", *Neuro computing,* Vol.396, pp.495-500, 2020

[26] Sharma, Radhika, Nishtha Goel, Nishita Aggarwal, Prajyot Kaur, and Chandra Prakash, "Next Word Prediction in Hindi Using Deep Learning Techniques", In 2019 *IEEE International Conference on Data Science and Engineering (ICDSE)*, pp.55-60, 2019

[27] Newsroom dataset, Available Online at: *https://huggingface.co/datasets/newsroom, Accessed On December, 2021*

[28] Cornell Movie–Dialogs Corpus (CMC),Available Online at: *https://metatext.io/datasets/cornell-movie--dialogs-corpus, Accessed On December, 2021*

Journal of Theoretical and Applied Information Technology
15th May 2022. Vol.100. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

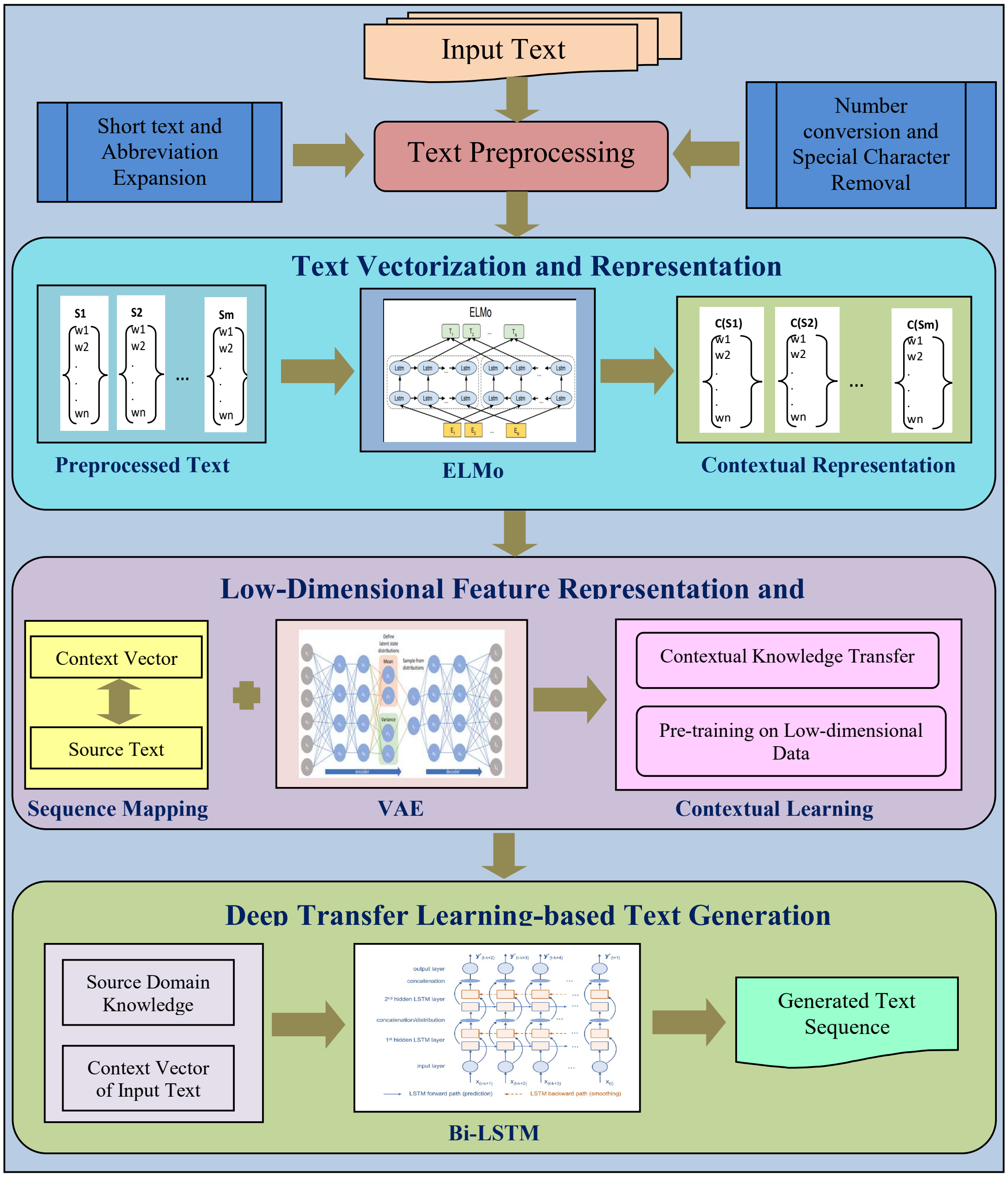


*Figure 1: An Illustration Of The Proposed DTGEN Methodology*

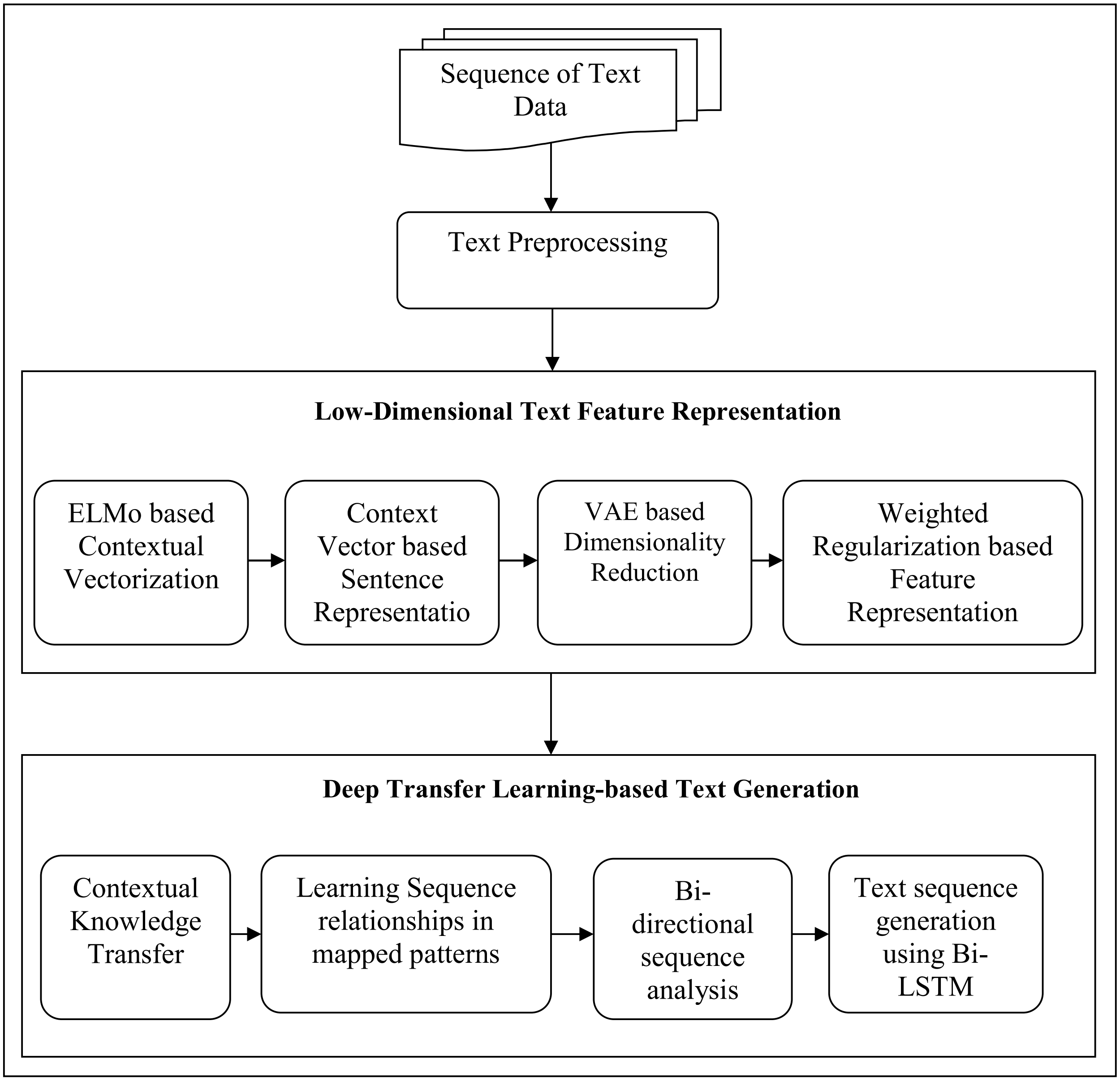


*Figure 2: The Proposed DTGEN Methodology*