A Systematic Literature Review on Text Generation Using Deep Neural Network Models

NOUREEN FATIMA, ALI SHARIQ IMRAN, (Member, IEEE), ZENUN KASTRATI, SHER MUHAMMAD DAUDPOTA, AND ABDULLAH SOOMRO

1Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan
2Department of Computer Science, Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway
3Department of Informatics, Linnaeus University, 35195 Växjö, Sweden

Corresponding author: Ali Shariq Imran (ali.imran@ntnu.no)

This work was supported in part by the Department of Computer Science (IDI), Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Gjøvik, Norway; and in part by the Curricula Development and Capacity Building in Applied Computer Science for Pakistani Higher Education Institutions (CONNECT) Project NORPART-2021/10502, funded by Diku.

ABSTRACT In recent years, significant progress has been made in text generation. The latest text generation models are revolutionizing the domain by generating human-like text. It has gained wide popularity recently in many domains like news, social networks, movie scriptwriting, and poetry composition, to name a few. The application of text generation in various fields has resulted in a lot of interest from the scientific community in this area. To the best of our knowledge, there is a lack of extensive review and an up-to-date body of knowledge of text generation deep learning models. Therefore, this survey aims to bring together all the relevant work in a systematic mapping study highlighting key contributions from various researchers over the years, focusing on the past, present, and future trends. In this work, we have identified 90 primary studies from 2015 to 2021 employing the PRISMA framework. We also identified research gaps that are further needed to be explored by the research community. In the end, we provide some future directions for researchers and guidelines for practitioners based on the findings of this review.

INDEX TERMS Systematic literature review, deep learning, text generation survey, natural language processing, quality metrics, neural network, GPT, LSTM.

I. INTRODUCTION

Text Generation is a field of study in Natural Language Processing (NLP) that combines computational linguistic and artificial intelligence to generate new text. It is a process of generating grammatically and semantically correct synthetic text. This process includes training a model that takes input data, learns the context from the input, and generates new text relating to the domain of input data. The generated text should satisfy the basic language structure and convey the desired message [1]. It is challenging to generate and evaluate grammatically, semantically, and synthetically correct text because text generation and its evaluation are open-ended. Thus, this Systematic Literature Review (SLR) discusses five research aspects associated with text generation. These include the deep learning approaches for text generation, quality metrics for evaluating generated text, training datasets used in the domain, languages on which the text generation is performed, and application areas for text generation.

Text generation can be performed at different granularity of text, i.e., character, word, and sentence level [2]. Text generation at a sentence level aims to analyze the entire text as a fine-grained and learn the relationship between the sentence and its context. Meanwhile, word-based text generation seeks to explore the structure of a sequence and predict the probability of the next word in a given text. Similarly, the model identifies the character rather than the entire document at character level text generation.

Automatic text generation was possible due to recent developments in computational resources coupled with advancements in deep learning techniques. Deep learning is a field of machine learning that uses artificial neural networks and representation learning. Text generation approaches can broadly
be categorized into three types of deep learning models as given below:

1) Vector-Sequence Model – Input is a fixed-size vector, whereas output can vary. For instance, this model can be used for caption generation of images [3].

2) Sequence-Vector Model - Input is of variable size, and output is a fixed-size vector. Classification is an example of this model [4].

3) Sequence-to-Sequence Model - Input and output are variable sizes in this model type. It is the most widely used variant of text generation models. Language translation belongs to this type of text generation model [5].

Above all, deep learning has contributed immensely to different aspects of natural language generation for various tasks including, dataset balancing [6], [7], next word prediction & text suggestion in chatting, generation of answers to questions in question answering system, in chatbots [8], [9], machine learning translation [10], [11], text summarization [12]–[14], text classification [15], [16], text generation for topic modeling [17], dialogue generation [18], sentiment analysis [19], [20], poetry writing [21], script writing for movies [1], [22], and others.

Evaluating the quality of the generated text governs the model’s performance and measures the diversity of generated and original text. The quality metrics are also known as evaluation methods. There are two ways to assess the quality of generated text: human-centric (HC) and machine-centric (MC) [23]. The human-centric evaluation method involves language and domain experts who evaluate the generated text. It is expensive in terms of both time and cost and is prone to human errors. On the other hand, the machine-centric evaluation method, as known as objective quality assessment, is widely adopted and found in the literature. It includes various evaluation metrics: Metric for evaluation of translation with explicit ordering (METEOR), bilingual evaluation understudy (BLEU), recall-oriented understudy for gesting evaluation (ROUGE), consensus-based image description evaluation (CIDER), National institute of standards and technology (NIST), word error rate (WER), Word Perplexity, and BERTScore. The machine-centric method saves time and cost, but the quality of an objective evaluation metric is highly language-specific.

There are various deep learning architectural frameworks widely used in the literature to implement deep learning models. Recurrent Neural Networks (RNN) [24] is one of them. It is a class of neural networks that uses the output of previous states as input in future states. This is the first algorithm that preserves the outputs of past states. One problem with RNN is that it forgets the previous outputs over a period of time due to a vanishing gradient. Bidirectional RNN [25] uses two RNN layers that look into the sequence in both directions, i.e., forward and backward, and combine their output. This is helpful when the current state is not only dependent on the previous state but also on the future state. One special class of RNN is Long Short-Term Memory (LSTM) [26] network that is used to retain the information of previous states over a very long period and forgets the irrelevant information. Gated Recurrent Unit (GRU) also overcomes the problem of vanishing gradient in RNN. GRU is a simplified version of LSTM.

Generative Adversarial Network (GAN) works on the concept of minmax game where the discriminator predicts if the sample is from the training set or is produced by a generative network, and the generator tries to maximize the mistakes of the discriminator.

GPT-2 was proposed by Radford et al. [27]. GPT-2 is a transformer-based model having 1.5 billion parameters. It is trained on 40GB of Internet text scrapped from eight million web pages. It is a revolutionary model in text processing. It has an exceptional human-like ability to generate long sequences.

In June 2020, OpenAI released the third version of GPT, which is 100 times larger than the previous model. GPT-3 is trained on 499 billion tokens of web data, and it has 175 billion parameters and 96 layers. It has more generative power that it may outperform in many different tasks like text generation, and zero-shot and one-shot learning [28]. However, the model is not publicly available; instead, API access was to be provided but to those who pay for it [29].

Usually, these pre-trained (LSTM and GPT-based) models are used to generate text in different domains. For example, it is possible to use the pre-trained GPT model for generating a movie script and customize its generation capability by fine-tuning using some movies’ script datasets. Once the data is gathered and model learning is customized to generate domain-specific text, the next step is to assess the quality of the generated text. LSTM was introduced as a character-level text generation model.

A. BACKGROUND

The basic strategy of text generation is first to train any language model on lots of sequences of text data, and then the model is capable of generating the next character or multiple characters in the sequence given previous characters as input. For example, generate the next character ‘k’ for given sequence ‘Cat likes mil’, as shown in Figure 1. Looking over the example mentioned in this paper [6], conditioning text, i.e., initial text that is fed to LSTM network for predicting next character in the sequence is ‘Cat likes m’. The LSTM model is trained on Wikipedia text or related domain of conditioning text. To predict the probability of the next character using the Softmax activation function at the output layer is defined in Equation 1.

\[
\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}
\]  

(1)

where, \(x_i\) is the LSTM score for character \(i\) to be the next character in conditioning text. Each of the \(x_i\) is not a probability score, therefore LSTM uses softmax to convert LSTM scores to probabilities score.

The actual magic of text generation is hidden in the sampling strategy. Text generation would be almost similar to
the original text if the next character is generated based on the highest score of probability taken from softmax output. Thus, some randomness was introduced in the generated text to introduce novelty and creativity to generated text. The sampling strategy introduces such randomness using temperature value. Suppose, $P_{original}$ is original probability distribution at Softmax, the $\alpha$ term is defined as,

$$\alpha = \log \left( \frac{P_{original}}{temperature} \right)$$

(2)

Once $\alpha$ is computed, the $P_{revised}$ is defined as,

$$P_{revised} = \frac{e^{\alpha}}{n}$$

(3)

where, $n$ is the number of elements in the original distribution and temperature value is an arbitrary value ranging from any non-zero value up to 1.

B. RELATED SURVEYS

There are a handful of surveys published on the topic of text generation, as shown in Table 1. We have found that five research papers have worked on a single aspect. Li et al. in [30] have provided a systematic literature review on deep learning approaches along with its data type. The authors mainly focused on the encoder and decoder-based deep learning architecture. Besides that, different data types (unstructured input, structured, and multimedia) were discussed, along with best-fitted transformer-based models. Similarly, Gatt et al., in [31], have worked on a systematic review of text generation-based applications. Lu et al. in [32], have provided a systematic literature review on the evolution matrix of SLR text generation. Lastly, both studies [23] and [33] have conducted a review study on text generation only based on quality metrics.

Besides that, few researchers have worked on two aspects of SLR text generation. Four research works [1], [31], [34], [36] have provided a literature review on quality metrics and deep learning approaches. Another review study focused on three aspects - quality metrics, approaches, and applications, and provided an overview of text generation [34]. Similarly, a study in [1] aimed to review multiple objectives of SLR on quality metrics, datasets, approaches, and application. There are two major limitations of these systematic literature reviews. First, seven articles were not peer-reviewed. Second, none of these attempts have worked on a comprehensive review of the dataset, quality metrics, languages, deep learning approaches, and trends in text generation in deep learning in a single study.

The limitations and findings shown in Table 1 provide a base for conducting a comprehensive review of text generation using deep learning. Therefore, our study focuses on articles published between 2015 to 2021. Ninety baseline articles are reviewed following the Preferred Reporting Items for Systematic Literature Review and Meta-Analysis (PRISMA) protocol for systematic literature review. We have investigated text generations on five different aspects, namely deep learning approach, quality metric, dataset, language, and application of text generation in deep learning, as shown in Figure 2.

The main contributions of this study are as follows:

1) A systematic map of 90 primary studies based on the PRISMA framework;
2) An analysis of the investigated text generations on five different aspects, namely deep learning approaches, quality metrics, datasets, languages, and applications on text generation in deep learning;
3) An overview of the challenges, opportunities, and recommendations of the field for future research exploration.

Additionally, this SLR provides an in-depth analysis, the most extensive and up-to-date body of knowledge of text generation based on five research aspects, and also focuses on the major challenges and future research directions in the text generation domain. To the best of our knowledge, there is no SLR on text generation that covers all these aspects.

The rest of the paper is organized as follows. Section II describes the research design of this SLR followed by Section III that covers the finding of RQs and provides the most relevant articles based on quality assessment criteria. Section IV provides the identified challenges and research gaps. Section V presents the recommendations and future research directions and finally Section VI summarizes the SLR.
II. RESEARCH DESIGN

In this study, we have applied systematic mapping as a research methodology for reviewing the literature. We have utilized the guidelines of PRISMA, given by [37]. This SLR consists of four major steps: planning and searching of primary studies, collection of studies, data extraction, and synthesis of data. The first step generally identifies research questions and objectives (stated in Section II-A). The search strategy step involves criteria for selecting studies, study selection procedure, keywords formulation for research and search queries, as well as the quality assessment criteria of extracted studies (which are addressed in Section II-B). The data extraction step involves strategies of data extraction from selected studies (see Section II-C and II-D for details). In addition, the final step involves Quality assessment (see Section II-E for more details).

A. RESEARCH QUESTIONS

The primary purpose of this SLR is to explore various techniques for text generation using deep learning. The following five research questions (RQs) were raised to achieve this aim, as shown in Table 2.

<table>
<thead>
<tr>
<th>RQ</th>
<th>Research question</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1</td>
<td>Which traditional and advanced deep learning approaches are used to generate text in the literature?</td>
</tr>
<tr>
<td>RQ2</td>
<td>What are the various metrics for evaluating the performance of text generation models?</td>
</tr>
<tr>
<td>RQ3</td>
<td>What are the major standard datasets for text generation in the literature?</td>
</tr>
<tr>
<td>RQ4</td>
<td>What are the application areas where text generation is extensively used?</td>
</tr>
<tr>
<td>RQ5</td>
<td>Which languages have been focused on text generation in deep learning?</td>
</tr>
</tbody>
</table>

B. RESEARCH OBJECTIVES

The following five research objectives of this study are given below:

- To investigate the existing traditional and advanced deep learning-based text generation approaches/techniques
- To explore various performance metrics used for evaluating text generation models
- To investigate various evaluation methods for measuring the quality of generated text
- To review the recent application domains where text generation is being applied
- To discuss the major challenges and future research directions in the text generation domain

C. SEARCH STRATEGY TO RETRIEVE PRIMARY STUDIES

The majority of studies have included text generation or automatic text generation as their data sources for text generation. Thus, various search keywords are formulated to retrieve the related literature from six reliable and high-quality academic databases, namely, Web of Science (WoS), Scopus, IEEE Xplore, Springer link, ScienceDirect, and ACM Digital Library. Five of the authors prepared a list of several relevant keywords to search the relevant literature on "text generation..."
techniques in deep learning” from the selected databases. Table 3 shows the keywords used to perform queries. Each keyword within the group is paired using the OR operator, whereas the groups are paired using the AND operator (see Table 3) to form a search query. The last row of Table 3 shows how keywords from different groups are concatenated to form a query that was executed in all six bibliographic databases.

The query was applied to the article title, article abstract, and article keywords to determine the relevant articles from the six selected bibliographic databases published in English from January 2015 to October 2021.

The search query identified 264 studies when applied to the five selected bibliographic databases, as shown in Figure 4. The identical studies from different databases were then extracted, and only distinctive copies were retained in EndNote for each primary sample. During removing of duplicate records, 50 studies were excluded.
The remaining 214 studies were analyzed after the removal of duplicate records. The screening was done based on the title, abstract, and keywords of the articles retrieved. These studies were retrieved by four authors using inclusion and exclusions criteria. A majority vote was used to include or remove articles for all inconsistencies. Furthermore, a final decision was taken in the event of ties between all the authors. Figure 4 indicates the screening of all the articles based on the title, abstract, and keyword-based screening method. Moreover, only 90 out of 264 were selected for primary studies; the remaining articles were excluded. The distribution of conference and journal reviewed papers are shown in Figure 5.

There were established criteria for excluding 117 articles. First, the purpose of many excluded studies was to extract information other the text generation. Second, the majority of the studies were about text classification, which is out of our scope. Third, a number of articles were written other than in English. Lastly, studies that were not peer-reviewed were excluded from the analysis thus to maintain the quality of this SLR paper.

We use the following inclusion criteria:
- The article must be used to include a generative model for text only
- The article must be published from 2015 to 2021
- The article must be published in a journal or a conference
- The article must be published in the English Language

We use following exclusion criteria:
- The articles which used NLP or machine learning techniques but did not propose or used any text generation techniques are excluded
- The articles published in languages other than English are excluded

E. QUALITY ASSESSMENT

The quality assessment criteria (QAC) were used to assess the quality of the 90 selected studies. The QAC was used to assess whether a selected primary study could achieve our review objectives. To determine the consistency of selected primary studies, a variety of questions were asked by all the authors. Table 4 describes the list of 10 questions to check the quality of studies. Either Yes or No can be the answer to each question with weights of 1 and 0 respectively. A group of four authors reviewed the selected primary studies. Results were evaluated after the quality assessment of each primary study. Finally, each question is matched by all the authors of the current research for every study for the review process. However, the quality review process did not rule out any study as all the studies fit the quality assessment questions. This review, therefore, included all the 90 studies selected.

III. SYSTEMATIC MAPPING STUDY RESULTS

In this section, we critically analyze 90 primary studies from five different aspects, namely, deep learning approach, quality metric, dataset, language, and application.

RQ1: Which traditional and advanced deep learning approaches are used to generate text in the literature?

There are two different approaches found in the literature to generate the text: traditional deep learning approaches (TDLA) and advanced deep learning approaches (ADLA). In traditional approaches, many deep learning-based models and NLP techniques were employed to generate the text. The topmost text generation models are RNN [38], LSTM [39], and CNN [40]–[42]. The text generation domain has seen some limitations due to its discrete nature [42]. Language generation requires a lot of effort, domain knowledge, and skills to learn the different semantic and contextual meanings from the text. Every language has its own standard rules and regulations. Therefore, it is not possible that the generated CNN for the English model may perform well in the Urdu language. Moreover, the contextual meaning of the generated text was a major issue in traditional text generation deep learning-based algorithms. Thus, to overcome the problem in traditional approaches, many advanced deep learning models were introduced like transformer [43], [44], BERT [45], GPT2 [46] and GPT3 [28]. These latest models are content-dependent algorithms with attention mechanisms [46]. Moreover, many approaches and models have been employed to generate text in different languages, which can be categorized into three main groups. Table 5 shows the papers grouped based on traditional approaches of text generation, advanced deep learning, and a combination of both approaches. 47 papers out of 90 have employed advanced approaches to generate the text, 40 papers used the traditional approaches, and 3 papers have used both approaches.

Moreover, we found that after 2018 there was a drastic increase in using traditional as well as advanced text generation approaches, as shown in Figure 6. Many researchers
have been working on text generation in various languages. The most often used algorithms for text generations in the reviewed studies for traditional approaches are LSTM and RNN. On the other hand, for advanced approaches, we have found different versions of GAN and BERT.

**RQ2: What are the various metrics for evaluating the performance of text generation models?**

Two metrics, human-centric and machine-centric, can measure generated text. In this SLR, we categorize studies into three groups based on the approaches used to assess the quality of the generated text. As shown in Table 6, 64 out of 90 papers have evaluated generated text on the basis of a machine-centric approach, 3 papers have evaluated on the basis of human experts, and 14 studies have utilized both the approaches human- and machine-centric. However, we found 9 studies that have not performed any measures to evaluate the generated text.

As shown in Figure 8 BLEU score has been widely used to check the quality of the generated text. 80% of studies used the BLEU score, 8% have used ROUGE and Perplexity, and 5% have used other metrics such as cosine similarity, content selection, diversity score, and word error rate.

**RQ3: What are the major standard datasets for text generation in the literature?**

The standard datasets for text generation based on their characteristic as mentioned below, have been extracted:

1) **Availability**: Private/Public
2) **Size**: Number of words, sentences, and reviews
3) **Type**: Sentence-, paragraph-, document-level and question/answer
4) **Format**: CSV, JSON, XML, files
5) **Annotation**: Labeled/unlabelled
6) **Quality**: Raw or pre-processed

The detail of datasets is given in Table 7. 9 out of 90 papers have used private datasets, and those are not publicly available [17], [49], [66], [67], [87], [89], [94]. 2 have used both public and private datasets [6], [84] and 79 studies have used the publicly available datasets, as shown in Table 7. In addition to this, 33 datasets were sentence-level, 14 were paragraph-level, 2 were document-level, 1 was question/answer type, and 1 study did not mention the type of dataset explicitly.

**RQ4: What are the application areas where text generation is extensively used?**

Text generation has wide range of applications in deep learning, which can be categorized into 18 groups as shown in Figure 9. We have found 10 out of 90 papers have main purpose to balance the dataset [6], [50], [68], [84], [85], [115], [118], [128], 8 out 90 papers have worked on data to text [47], [55], [64], [82], [82], [100], [102], [108], and speech to text [67], [83], [98], [103]–[105], [110], [113], respectively. 7 papers have worked on script writing [3], [17], [57], [60], [85], 5 papers have worked on machine translation [10], [11], [56], [87], [101], [120]. Apart from these, 4 papers have worked on text summarization [49], [87], [97], [126] and 2 papers [1], [90] have worked on abstract meaning representation (AMR)- AMR to text goal is to generate sentences from abstract meaning representation graphs and its seq2seq or graph2seq problem. In addition, we found 2 applications of product reviews [129], [129] of mobile devices that have worked in text generation. The single paper was found on C programming language code generation [106] and online shop guideline generation [84].

**RQ5: Which languages have been focused on text generation in deep learning?**

Many text generation models and approaches have been employed to generate text in different languages. There are eleven languages found in the literature: English,
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Size</th>
<th>Quality</th>
<th>Format</th>
<th>Lab</th>
<th>Private/Public</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford Natural Language Intemnet (SNLI)</td>
<td>Sentence</td>
<td>30,000</td>
<td>Pre-process</td>
<td>Json</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://nlp.stanford.edu">http://nlp.stanford.edu</a> SNLI/</td>
</tr>
<tr>
<td>VQA2 Task</td>
<td>Paragraph</td>
<td>1,285,780 instances</td>
<td>Raw</td>
<td>txt</td>
<td>N</td>
<td>N</td>
<td><a href="http://image-net.org/">http://image-net.org/</a></td>
</tr>
<tr>
<td>HARE: Arabic Reviews in Arabic Dataset</td>
<td>Paragraph</td>
<td>50,000 books reviews</td>
<td>Pre-process</td>
<td>txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://github.com/HARE-Arabic-Dataset">http://github.com/HARE-Arabic-Dataset</a></td>
</tr>
<tr>
<td>AI4MultiLaner</td>
<td>Paragraph</td>
<td>No mentioned</td>
<td>Pre-process</td>
<td>txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.aimm.org/datasets/AI4MultiLaner">http://www.aimm.org/datasets/AI4MultiLaner</a></td>
</tr>
<tr>
<td>Arabic-Review Project Corpus (ARPC)</td>
<td>Sentence</td>
<td>141,8 million reviews</td>
<td>Pre-process</td>
<td>txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://nlp-db.loni.ucla.edu/">http://nlp-db.loni.ucla.edu/</a></td>
</tr>
<tr>
<td>Arabic-Chinese-English-Shopping Corpus</td>
<td>Sentence</td>
<td>150 million words</td>
<td>Not mentioned</td>
<td>Not mentioned</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://aima.ucd.ie/">http://aima.ucd.ie/</a></td>
</tr>
<tr>
<td>Yelp review dataset</td>
<td>Sentence</td>
<td>Not mentioned</td>
<td>Pre-process</td>
<td>Json</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Sentence token base dataset</td>
<td>Paragraph</td>
<td>Not mentioned</td>
<td>Pre-process</td>
<td>txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>NEWS dataset NYT/cnn</td>
<td>Paragraph</td>
<td>579,795 articles</td>
<td>Pre-process</td>
<td>txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.bag-of-words.uk/">http://www.bag-of-words.uk/</a></td>
</tr>
<tr>
<td>IMDB corpus (LDC2017T05)</td>
<td>Sentence</td>
<td>50,235 sentences</td>
<td>Raw</td>
<td>XML</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.lrec-conf.org/">http://www.lrec-conf.org/</a></td>
</tr>
<tr>
<td>Chinese-En-Unsupervised platform, Tencent</td>
<td>Sentence</td>
<td>Not mentioned</td>
<td>Not mentioned</td>
<td>Not mentioned</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://nlp.benuo.edu.cn/">http://nlp.benuo.edu.cn/</a> dataset.html</td>
</tr>
<tr>
<td>Amazon Product Reviews</td>
<td>Sentence</td>
<td>141,8 million reviews</td>
<td>Pre-process</td>
<td>txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.bag-of-words.uk/">http://www.bag-of-words.uk/</a></td>
</tr>
<tr>
<td>IMDb Movie Reviews</td>
<td>Sentence</td>
<td>Not mentioned</td>
<td>Pre-process</td>
<td>csv, txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Arabpoetry</td>
<td>Sentence</td>
<td>8,000 Poems</td>
<td>Pre-process</td>
<td>csv</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>kagispanish</td>
<td>Paragraph</td>
<td>1 million words</td>
<td>Raw</td>
<td>Not mentioned</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Twitter product reviews</td>
<td>Paragraph</td>
<td>4,235,258 reviews</td>
<td>Pre-process</td>
<td>csv</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>MacauDStorytelling</td>
<td>Paragraph</td>
<td>Not mentioned</td>
<td>Pre-process</td>
<td>txt</td>
<td>N</td>
<td>N</td>
<td>Public</td>
</tr>
<tr>
<td>Amazon Review</td>
<td>Sentence</td>
<td>270,000 sentences</td>
<td>Pre-process</td>
<td>txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.nanosad.com/">http://www.nanosad.com/</a></td>
</tr>
<tr>
<td>COCO Image Captions</td>
<td>Sentence</td>
<td>20,734 words and 417,136 sentences</td>
<td>Pre-process</td>
<td>Not mentioned</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.vision.ee.ethz.ch/">http://www.vision.ee.ethz.ch/</a></td>
</tr>
<tr>
<td>Chinese Poems</td>
<td>Sentence</td>
<td>Not mentioned</td>
<td>Not mentioned</td>
<td>Not mentioned</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.hypergrid-brain.com/">http://www.hypergrid-brain.com/</a></td>
</tr>
<tr>
<td>ROCTWORY</td>
<td>Paragraph</td>
<td>4950 sentences</td>
<td>Pre-process</td>
<td>csv</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Brown Corpus</td>
<td>Paragraph</td>
<td>10 million words</td>
<td>Pre-process</td>
<td>Json</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Penn Treebank</td>
<td>Sentence</td>
<td>90,000 words</td>
<td>Pre-process</td>
<td>txt</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>ROYOWRES</td>
<td>Document</td>
<td>4853 narratives</td>
<td>Pre-process</td>
<td>text</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Sentence</td>
<td>738,321 biographies</td>
<td>Pre-process</td>
<td>txt and csv</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.bag-of-words.uk/">http://www.bag-of-words.uk/</a></td>
</tr>
<tr>
<td>Student Reviews</td>
<td>Sentence</td>
<td>5000 reviews</td>
<td>Pre-process</td>
<td>Not mentioned</td>
<td>N</td>
<td>N</td>
<td>Public</td>
</tr>
<tr>
<td>Tweet review</td>
<td>Sentence</td>
<td>Not mentioned</td>
<td>Pre-process</td>
<td>Json</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>MOOC Lecture Transcripts</td>
<td>Sentence</td>
<td>878 thousand sentences</td>
<td>Pre-process</td>
<td>csv</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Large Movie Review</td>
<td>Sentence</td>
<td>5000 reviews</td>
<td>Pre-process</td>
<td>csv</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Obama Speech</td>
<td>Sentence</td>
<td>Not mentioned</td>
<td>Pre-process</td>
<td>Json, MP3</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>MapLab WEHNO</td>
<td>Paragraph</td>
<td>21,835 dataset</td>
<td>Raw</td>
<td>XML, Json</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Restaurant Dataset</td>
<td>Sentence</td>
<td>Not mentioned</td>
<td>Pre-process</td>
<td>csv</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Finding a hotel, buying a laptop</td>
<td>Sentence</td>
<td>Not Mentioned</td>
<td>Raw</td>
<td>dat</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>MECOCO</td>
<td>Question/Answer</td>
<td>1.7 million Q/A pairs</td>
<td>Pre-process</td>
<td>Json</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>ROBOCUP</td>
<td>Sentence</td>
<td>650 sentences</td>
<td>Pre-process</td>
<td>txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Kaggle Arabic Poems</td>
<td>Sentence</td>
<td>50k Poems</td>
<td>Pre-process</td>
<td>csv</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>ULSS days</td>
<td>Sentence</td>
<td>201,189 headlines</td>
<td>Pre-process</td>
<td>csv</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Amazon Review Food reviews</td>
<td>Sentence</td>
<td>500,000 food reviews</td>
<td>Pre-process</td>
<td>csv</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Hotel Dataset</td>
<td>Sentence</td>
<td>Not Mentioned</td>
<td>Pre-process</td>
<td>txt</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Google SearchesコンテンツScoring/25</td>
<td>Sentence</td>
<td>10,000 sentence</td>
<td>Pre-process</td>
<td>Json</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Cornell Movie Dialog</td>
<td>Sentence</td>
<td>50,713 utterances</td>
<td>Pre-process</td>
<td>Json</td>
<td>N</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Restaurants</td>
<td>Sentence</td>
<td>11.5k reviews</td>
<td>Pre-process</td>
<td>csv</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>Data for parliament speeches</td>
<td>Sentence</td>
<td>12,000</td>
<td>Pre-process</td>
<td>Not mentioned</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>sentimentSLU</td>
<td>Paragraph</td>
<td>19 FE words</td>
<td>Raw</td>
<td>Json</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
<tr>
<td>LOGICALS</td>
<td>Sentence</td>
<td>122K words</td>
<td>Raw</td>
<td>Json</td>
<td>Y</td>
<td>Public</td>
<td><a href="http://www.kaggle.com/">http://www.kaggle.com/</a></td>
</tr>
</tbody>
</table>

Chinese, Bengali, Arabic, Russian, Korean, Slovak, Spanish, Czech, German, and Macedonian. This information can help researchers which languages lack research in this domain, which languages need to have focused more on, and what possible deep learning approaches could contribute to a specific language. As can be seen from Figure 7, 74% of the articles have worked in the English language, 7% of articles have worked in Bengali, 2% of the articles generated Arabic and Russian, and 1% of the articles found for rest of languages. Moreover, the detailed summary of languages according to deep learning approaches is shown in Figure 10.

Many resources are available for English and Chinese languages like Dataset, lexical, syntactic, and POS tagging and programming development support. Therefore, both languages are known as rich-resource languages. On the other hand, languages such as Bengali, Arabic, Russian, Korean, Slovak, Spanish, Czech, German, and Macedonian are known as low-resource languages because resource availability is scarce.
A brief summary of language on the basis of deep learning techniques is depicted in Table 8, Table 9 and in Figure 10. A variety of traditional and advanced deep learning-based approaches have been used for English text generation, as shown in Table 9. Yet, it needs more experiments with text generation in these languages like Turkish, Hindi, Russian, Macedonian, German, Czech, Spanish, Slovak, Korean, Bengali, and Arabic, as shown in Table 8. In addition to this, a brief summary of language year-wise is depicted in Table 8. The research work on English text generation has fast-growing after 2018 and found 74 studies. After English, we found that studies on Chinese language text generation are constantly growing and found 6 studies. A huge void is left for the other languages to benefit from in this domain.

### IV. IDENTIFIED GAPS

In this section, we discuss the major gaps in some areas concerning text generations on the basis of five aspects that need further research and development. The following list shows some gaps which are mapped on our research questions, as shown in Table 10.

1) **Complex language constructs.** Language construct is a piece of language syntax, and every language has its own language constructs. Therefore, it may vary from language to language. For example, the language construct for English sentence is **Subject + Verb + Object**, whereas for Urdu language it is **Subject + Object + Verb**. There is no proper way to deal with complex language that requires construct morphology, delexicalised verbs, and abbreviations. Many researchers currently adopt a translation method where low-resource language is being converted to English. However, it mainly provides rigid word order and relatively poor morphology because the technique developed for English may not work for other low-resource languages [130].

2) **The demand for diversity.** Most of the generated text approaches found in the studies discussed in this SLR have a redundant and poor quality text generation problem [21], [49].

3) **Improper selection of quality metrics.** We observed in this survey paper that the selection of quality metrics was improperly employed in many of the studies.

---

**TABLE 9. Text generation approaches for english language.**

<table>
<thead>
<tr>
<th>Language</th>
<th>Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>RNN, GRU, LSTM, Bi-RNN, Bi-LSTM (Encoder and Decoder), RNN, GRU, LSTM, Bi-RNN, Bi-LSTM (attention), SeqGAN, TranGAN, Rel-GAN, PAN, VAE</td>
</tr>
</tbody>
</table>

**TABLE 10. Gaps linked to research question.**

<table>
<thead>
<tr>
<th>Identified Gaps</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex language constructs</td>
<td>RQ1</td>
</tr>
<tr>
<td>The demand for diversity</td>
<td>RQ1</td>
</tr>
<tr>
<td>Improper selection of quality metrics</td>
<td>RQ2</td>
</tr>
<tr>
<td>Limited resources</td>
<td>RQ3</td>
</tr>
<tr>
<td>Scarcity of datasets</td>
<td>RQ3</td>
</tr>
<tr>
<td>Un-standardized source of datasets</td>
<td>RQ3</td>
</tr>
</tbody>
</table>
found in the literature, and the quality of the generated text was not properly evaluated. For example, the BLEU metric is used to measure the quality of two sentences; thus, it works well for the short sentence-based problem. However, it may not capture the semantic meaning and does not map well to human judgmental capacity. Keeping in view this point, AMR is a seq2seq generation or graph to sequence generation. It is known for semantics. However, many authors have validated the quality of generated text by using BLEU metrics [90].

4) **Limited resources.** Lack of resources such as dictionaries and POS taggers for low-resource languages such as Bengali, Arabic, Russian, Korean, Slovak, Spanish, Czech, German, Urdu, Hindi, Macedonian, etc. There are thirty language published dependency tree-bank reported in [131]. These languages have been working with Google Translate (support 80 different languages), as for the majority of languages, there is no support for NLP resources at all [132].

5) **Scarcity of datasets.** Lack of benchmark datasets for low-resource languages, including Bengali, Arabic, Russian, Korean, Slovak, Spanish, Czech, German, Urdu, Hindi, Macedonian, etc. For instance, we have found the three studies for Bengali text generation [87], [94], [133], and all these studies have used their own extracted dataset. The major reason for using their dataset was the lack of publicly available datasets for the Bengali language. Similarly, there is no dataset available for the Czech language; we have found only one paper for the Czech language in which a multilingual dataset is used [119].

6) **Un-Standardized sources of dataset.** We found that a wide variety of sources of datasets are available, like Quora, GitHub, Kaggle, own website. Sometimes, there is a considerable amount of noise available in datasets; therefore, researchers need to adopt a lot of prepossessing techniques to get the best results from the noisy datasets [134].

### V. RECOMMENDATIONS AND FUTURE RESEARCH DIRECTIONS

In this section, we highlight various research directions for researchers in the field, which require considerable efforts to improve the performance of the text generation domain. These research directions are presented below.

#### A. STANDARDIZED DATASET

Research work is required to develop a few benchmark datasets for Arabic, Chinese, Bengali, Russian, Korean, Slovak, Spanish, Czech, German, Macedonian, and other low-resource languages. The Standardized dataset formation can be at the document level, question/answer form, and paragraph level. In addition to this, we recommend that researchers explore these benchmark datasets: Books3 Stack Exchange, PubMed Abstracts, and CC-2021-04 for various text generation applications, including automatic article summarization and generating synthetic samples to deal with the data imbalance problem.

#### B. QUALITY METRICS

Our study showed that researchers evaluated generated text using machine and human-based approaches. Nonetheless, a considerable number of research articles failed to evaluate the quality of generated text [17], [40], [49], [87], [115], [133], although they reported excellent results. These results may be biased, in which the experiments that obtained low results may not have been reported. To deal with this issue, we recommend a standard way to evaluate the generated text depending on the nature of the generated text. For instance, for text summarization, ROUGE quality metric is recommended.

Another issue found in the literature is the selection of an inappropriate metric for text generation quality assessment. For example, the BLEU metric is used to measure the quality of two sentences; thus, for a short sentence-based problem, it works well. However, it may not capture the semantic meaning and does not map well to human judgmental capacity. Keeping in view this point, AMR is a seq2seq generation or graph to sequence generation. It is known for semantics. However, many authors have validated the quality of generated text by using the BLEU metrics.

#### C. TEXT GENERATION IN LOW-RESOURCE LANGUAGES

We have observed a high demand and scope for text generation in low-resource languages in this SLR. A majority of studies have worked on English language text generation. Nonetheless, we have found that 23% of researchers have worked on low-resource languages. Low-resource languages such as Arabic, Spanish, Turkish, Slovak, Hindi, Russian, Macedonian, Czech, Bengali, Korean, Urdu, and alike require significant efforts in this domain. There exist loads of online text thanks to social media and news websites that such languages can benefit from in training language models for text generation. Thus, future research may explore and benefit from available resources for text generation through deep neural network models. Moreover, advanced deep learning approaches for text generation such as GPT-2, BERT, and ELMo should be further considered for exploration by researchers in this field as they have outperformed other methods in the English language [6].

#### D. USE OF GPT-3 FOR TEXT GENERATION

Existing studies either used traditional or advanced deep learning approaches for text generation. Thus, researchers can emphasize generating text using GPT-3, which is trained on 499 Billion tokens of web data and has 175 billion parameters and 96 layers. It has greater ability and generative power that it may outperform other algorithms in many different tasks like text generation [28].
E. NLP BASIC OPERATIONS IN LOW-RESOURCE LANGUAGES

Standard NLP operations like POS-tagging, tokenization, lemmatization, stemming, word meaning, and related tasks are extremely important in ensuring the quality of the generated text. In low-resource language, there exists an enormous scarcity of these standard basic tasks. Researchers are highly encouraged to come forward and contribute in these areas to further democratize the Internet with increased use of local languages alongside English and other resource-rich languages. It is essential to mention here that loads of mature algorithms are available in the field for these NLP tasks. Data availability is also not an issue. Only a few concentrated efforts are required to work on basic NLP tasks in low-resource languages. These efforts would certainly promote low-resource languages on the Internet.

VI. CONCLUSION

Text generation is the creative side of AI. For decades, computer scientists have been promising humanity to bring artificial intelligence equal to artificial general intelligence. We have to ensure AI can generate text that can pass the Turing test to fulfill these promises. In the past few years, we only recently observed that our dream of synthetic text generation is very close to reality, albeit it is only for a few resource-rich languages.

Text generation has gained wide popularity because a profusion of applications uses them, and there is abundant availability of text online thanks to social media, news outlets, and other sources with enormous usage of text. A few applications which are benefited from text generation include generating and predicting character/word/sentence while typing an email or chatting, chatbot, movie/drama scriptwriting, poetry generation, and many other applications. Moreover, text generation has also been attracting the attention of researchers in the application area of education, industry, and social networks to provide an insight view on different aspects of the approaches. In this context, this systematic literature review provides an analysis of the investigated 90 relevant papers (2015 to 2021) based on text generation in five different aspects; namely text generation approaches, quality metrics, dataset, languages, and applications on text generation in deep learning.

After thoroughly mapping the primary articles, we reviewed them critically to explore different aspects of text generation. For instance, diverse quality metrics are applied to evaluate the generated text. A myriad of approaches is proposed for text generation. A variety of datasets exist; our review concerned their size, format, and applications in which text generation has been applied.

We have provided an overall trend of publications investigating deep learning approaches for text generation throughout the studied years. We have noticed that there is a significant growth of articles published during the year 2018, where the advanced deep learning techniques were mostly represented. In addition to this, text generation in the English language has been more exploited in literature than in any other language.

This systematic literature review will help researchers, academicians, practitioners, and educators who are interested in text generation with data sources, approaches, trends, techniques, and languages.

REFERENCES


N. Fatima et al.: Systematic Literature Review on Text Generation Using Deep Neural Network Models
REFERENCES


ALI SHARIQ IMRAN (Member, IEEE) received the master’s degree in software engineering and computing from the National University of Sciences and Technology (NUST), Pakistan, in 2008, and the Ph.D. degree in computer science from the University of Oslo (UiO), Norway, in 2013. He is currently associated with the Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway, as an Associate Professor. He specializes in applied research with a focus on deep learning technology and its application to signal processing, natural language processing, and semantic web. He has over 65 peer-reviewed journals and conference publications to his name. He is a member of the Norwegian Colour and Visual Computing Laboratory (Colourlab) and the Norway Section. He served as a Reviewer for many reputed journals over the years, including IEEE Access, as an Associate Editor.

ZENUN KASTRAI received the master’s degrees in computer science through the EU TEMPUS Program developed and implemented jointly from the University of Pristina, Kosovo, the Université de La Rochelle, France, and the Institute of Technology Carlow, Ireland, and the Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Norway, in 2018. He is associated with the Department of Informatics, Linnaeus University, Sweden. He is the author of more than 40 peer-reviewed journals and conferences. His research interests include artificial intelligence with a special focus on NLP, machine/deep learning, and sentiment analysis. He has served as a reviewer for many reputed journals over the years.

SHER MUHAMMAD DAUDPOTA received the master’s and Ph.D. degrees from the Asian Institute of Technology, Thailand, in 2008 and 2012, respectively. He is currently working as a Professor in computer science with Sukkur IBA University, Pakistan. Alongside his computer science contribution, he is also a Quality Assurance Expert in higher education. He has reviewed more than 50 universities in Pakistan for quality assurance on behalf of Higher Education Commission in the role of educational quality reviewer. He is the author of more than 35 peer-reviewed journals and conferences. His research interests include deep learning, natural language processing, and video and image processing.

NOUREEN FATIMA received the master’s degree in computer science from Sukkur IBA University, Sukkur, Pakistan, in 2021. She is currently a Research Assistant with the Center of Excellence for Robotics, Artificial Intelligence, and Blockchain, Department of Computer Science, Sukkur IBA University. She is the author of several articles published in international journals. Her research interests include applied research in the field of artificial intelligence and its application to signal processing and natural language processing. She served as a Reviewer for IEEE Access.

VOLUME 10, 2022 53503