IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 8, 2022, accepted February 2, 2022, date of publication February 7, 2022, date of current version February 16, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3149798

Framework for Deep Learning-Based Language Models Using Multi-Task Learning in Natural Language Understanding: A Systematic Literature Review and Future Directions

RAHUL MANOHAR SAMANT[1], MRINAL R. BACHUTE[2], SHILPA GITE[3], AND KETAN KOTECHA[3]

[1]Computer Science and Information Technology Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2]Department of Electronics and Telecommunication Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[3]Symbiosis Centre of Applied AI (SCAAI), Symbiosis International (Deemed University), Pune 412115, India

Corresponding author: Mrinal R. Bachute (mrinal.bachute@sitpune.edu.in)

ABSTRACT Learning human languages is a difficult task for a computer. However, Deep Learning (DL) techniques have enhanced performance significantly for almost all-natural language processing (NLP) tasks. Unfortunately, these models cannot be generalized for all the NLP tasks with similar performance. NLU (Natural Language Understanding) is a subset of NLP including tasks, like machine translation, dialogue-based systems, natural language inference, text entailment, sentiment analysis, etc. The advancement in the field of NLU is the collective performance enhancement in all these tasks. Even though MTL (Multi-task Learning) was introduced before Deep Learning, it has gained significant attention in the past years. This paper aims to identify, investigate, and analyze various language models used in NLU and NLP to find directions for future research. The Systematic Literature Review (SLR) is prepared using the literature search guidelines proposed by Kitchenham and Charters on various language models between 2011 and 2021. This SLR points out that the unsupervised learning method-based language models show potential performance improvement. However, they face the challenge of designing the general-purpose framework for the language model, which will improve the performance of multi-task NLU and the generalized representation of knowledge. Combining these approaches may result in a more efficient and robust multi-task NLU. This SLR proposes building steps for a conceptual framework to achieve goals of enhancing the performance of language models in the field of NLU.

INDEX TERMS Deep learning, knowledge representation, multi-task NLU, unsupervised learning.

I. INTRODUCTION

NLU is a relatively new research topic in which a computer analyses and extracts information from natural language text before doing standard NLU tasks, viz. information retrieval, question-answering, language translation, text summarization, news classification, and so on. The recent trends in text mining caters to the ever-increasing need of extraction of high-quality information from structured as well unstructured text. On the contrary, the recent trends in systematic language understanding (SLU) are in the direction of understanding actionable intents in the input text, along with grammatical structural correctness of the input language. The application domains for NLU are listed in Table 1. The popular business application based on NLU is a chatbot. According to Gartner's AI customer service statistics [1], chatbots will be responsible for 85% of customer service by 2020. According to Crunchbase's AI stats [1], more than 10,000 developers now build chatbots for Facebook Messenger.

According to Juniper statistics [2], Chatbots are expected to cut business costs by $8 billion. These statistics predict the emergence of AI-powered chatbots (conversational AI) in

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

TABLE 1. Application domains for NLU.

Domain	Applications
Machine translation	IBM Watson
Task-based dialogue-based systems	Booking tickets, taking an appointment using Google assistant
Large scale content analysis	Summarizers
Text categorization	Sarcasm detection, sentiment analysis
Voice-activation	Alexa, Cortana

many business sectors like banking, education, tourism, legal, and government, where customer interaction and customer experience can be designed using AI-based techniques.

A. RELEVANCE AND SIGNIFICANCE

Because the general goals of AI are to make computers and intelligent devices listen, talk, and understand language; think and solve problems; and create new things, research in the field of NLU is relevant to many aspects of AI. Most NLU's tasks entail reading, interpreting, and categorizing material. This requirement necessitates the creation of systems capable of answering questions after reading a paragraph or document. It necessitates human-like language comprehension abilities. Machine reading comprehension can also be employed in virtual assistants so that after reading documents, these assistants can aid in answering customer service concerns. It can also be utilized in the workplace to assist users in reading and processing emails or large-scale business papers, as well as summarizing pertinent information. In-home automation also voice-activated assistants help users communicate with various home appliances in meaningful ways.

B. EVOLUTION OF DEEP LEARNING MODELS USED IN NLU TASKS

The underlying major sub-task in all NLU tasks is text classification or categorization (TC). Text data belongs to heterogeneous sources like social media, electronic communication, or interrogative data like QA from the client interaction. Text is an excellent basis of the information, but inferring useful information can be complex and laborious due to its unstructured style. Text categorization (TC) can be achieved through manual or automatic labelling. Due to the availability of explosive data in text form in many applications, automated text categorization is becoming one of the most effective methods. There are two main categories of Automatic text classification - rule-based and AI-based methods. The first category of Rule-based methods group text into different classes using a set of predefined rules and require a deep knowledge about related domains. The second type, AI-based approaches, learn to classify text based on the training of data using pre-labelled examples. An ML algorithm learns the relation between the text and its labels. Most classical machine learning-based models follow the staged procedure. The majority of NLU's tasks entail reading, comprehending, and interpreting. Some features are manually retrieved from any document in the first phase, and those features are then fitted to a classifier to create a prediction in the second step. A bag of words (BoW) is an example of a manually extracted feature, and Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF) are prominent classification techniques. This staged approach has several limitations; for example, depending on the manually extracted features requires complex feature analysis to obtain decent performance. Due to its considerable dependence on domain local knowledge for features engineering, the generalization of this method for new tasks demands more effort. Further, these models cannot exploit the availability of colossal training data because of the predefined features. Neural Network methodologies have been used to overcome the restrictions by using manually extracted features. The main focus of these approaches depends on a machine learning algorithm that maps text into a low dimensional continuous feature vector, so manually extracted features are not needed. Deerwester *et al.* [3], in 1989, proposed Latent Semantic Analysis (LSA). It is one of the earliest embedding models. LSA is a linear model with fewer than one million parameters that have been trained on the corpus of 200K words. Some of the limitations of this model are the statistical properties of an LSA space when used as a measure of similarity and the limited use of dimensional information in the vector representation. Bengeo *et al.* [4] introduced the first natural language model in 2000. It is based on a feed-forward neural network and trained using 14 million words. These premature embedding models, on the other hand, underperform traditional models based on manually derived features. This scenario drastically changed when much larger embedding models with much larger training data were developed. In 2013, Google developed a series of word2vec models [5] trained using the corpus of 6 billion words and immediately became state of the art for many NLU tasks. In 2017, AI2 and the University of Washington collaborated to create a contextual embedding model based on a three-layer bidirectional LSTM using 93 million parameters and 1 billion words. Because it captures contextual information, the Elmo [6] model performs substantially better than the word2vec approach. In 2018, OpenAI began developing embedding models using Google's Transformer [7], a revolutionary neural network architecture. The transformer is entirely dependent on attention, which enhances the accuracy of large-scale model training on TPU significantly. GPT [8], the first model designed with Transformers, is currently commonly utilized for text generation jobs. Google also developed BERT [9] based on the bidirectional transformer in 2018. BERT consists of 340M parameters and is trained using a corpus of 3.3 billion words. The trend of using bigger models with more training data continues with the recent introduction of OpenAI's latest GPT-3 model [10]. It has 170 billion parameters, and Google's Gshard [11] contains 600 billion parameters. Other popular models based on generative pre-trained transformer techniques include T-NLG from Microsoft with 17 billion parameters and Megatron

from NVIDIA with 1 trillion training parameters. [12] Although these massive-scale models perform admirably on certain NLU tasks, other researchers contend they lack language understanding and are unsuitable for many mission-critical applications. [13], [14]. The evolution of deep learning models in NLP and NLU is depicted in Figure 1

II. PRIOR RESEARCH

One of the objectives of this SLR is to explore the existing deep learning models in the area of NLU for multi-tasks. Figure 2 outlines various sections in this paper.

There is a scarcity of SLRs in the research topic of NLU. The review [15] is one of the recent and pertinent surveys on deep learning models using transformers as an underlying architecture. This review focus on knowledge encoding techniques for transformer-based models. It also points out the challenges faced by these models as context and language dependence issues. The highlight of this study is establishing BERT as the backend model in all such variants. Another study [15] provides an extensive review of multi-task learning. (MTL) The study contributes to the domain by offering practical approaches into settings of MTL that are introduced for supervised learning, unsupervised learning, semi-supervised, active learning, reinforcement learning, online learning, and multi-view learning. It also suggests parallel and distributed MTL for improving the speed and performance of those models. It also presented recent theoretical analyses for MTL. The survey [16] summarized and examined the current state-of-the-art (SOTA) NLP models for standard NLP tasks for optimal performance and efficiency. The significant contribution of this survey is to provide a detailed understanding and functioning of the different architectures, a taxonomy of NLP, NLU, and NLG designs, and comparative evaluations. This survey [17] rightly pointed out that the self-attention mechanism and transformer-based architectures exponentially improve the performance of language models.

There are a few constraints of the previous research, which is enlisted as follows:

1. Current surveys are task-specific and architecture-centered.
2. Current literature does not check the generalization of language models to be suitable for all NLU tasks.
3. Very few surveys discuss the knowledge representation methods for multi-tasks.
4. Few surveys examine the existing online tools to build a general-purpose framework for multi-tasks NLU.

This SLR is comprehensive in terms of examining the current trends and challenges related to building a general-purpose framework for multi-task NLU, and quality of available benchmarking datasets in the public domain, and the techniques used for creating such a framework.

A. MOTIVATION

There is no prevailing SLR with the exhaustive examination of general-purpose language models for multi-task NLU covering explicit benefits, comparative analysis, taxonomies, and pit-falls. Table 2 lists research questions.

TABLE 2. Research questions.

RQ1. What are the different algorithmic approaches available in combining MTL to improve the accuracy of the proposed framework for NLU?
RQ2. Which are the standard benchmarking datasets for MTL- NLU tasks evaluation?
RQ3. Which are the learning methods for improving the learning performance of NLU in combining multiple tasks?
RQ4. Which techniques are effective for reducing the need for huge annotated data samples?
RQ5. Which are prevalent knowledge representation techniques for MTL- NLU?

B. GOALS FOR THIS RESEARCH

NLU mainly comprises tasks like inference, text entailment, sentiment analysis and named entity recognition. The field of NLU research aims to attain the task proficiency for these tasks contained in standard benchmarking datasets like GLUE (General Language Understanding Evaluation) and superGLUE. (Super GLUE). The goal of this research is to first match and then surpass the established score of existing models in the literature. This SLR aims at recognizing and judgmentally examining the research papers and their output concerning the framed research questions. The RQs of interest are listed in Table 2.

C. CONTRIBUTION OF THIS STUDY

The contributions of this SLR:

1. This study identified 93 primary studies on language models in NLU from 2011 to 2021.
2. A detailed study of benchmarking datasets in the public domain is made. A suitable benchmark for a general-purpose framework of language models for multi-task NLU is provided.
3. A summary of available online tools for building a general-purpose framework of language models for multi-task NLU is presented.
4. The research gaps were identified. These gaps lead to future directions in the research area of NLU.
5. A conceptualizing framework for the general-purpose language models with enhanced transformer encoding with active learning for multi-tasking NLU is proposed to produce this SLR.

III. METHODOLOGY FOR RESEARCH

The guiding principles introduced by Kitchenham and Charters [18] were followed for preparing this SLR. Table 5 depicts the techniques of PIOC (Population, Intervention, Outcome, Context) utilized for enclosing the research questions. The procedural flowchart for this process is shown in Figure 3.

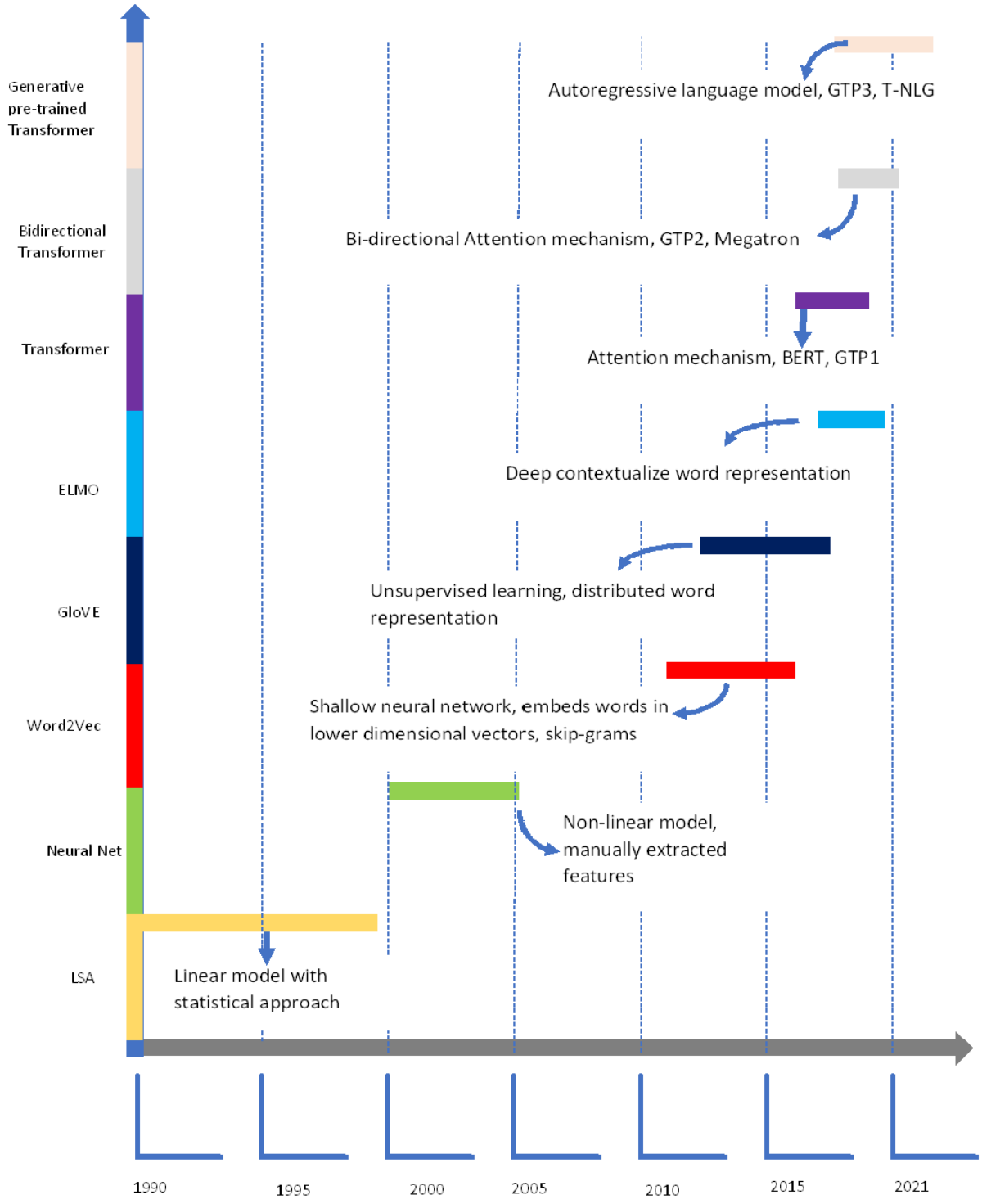


FIGURE 1. Evolution of deep learning models in NLU.

A. RESEARCH STUDIES SELECTION CRITERIA

The key phrases were selected to acquire the required search results to inquire about the RQs of the domain. The search string is shown below:

(multi-task nlu'' OR ''multi-task nlu framework'' OR'' natural language understanding'' AND ''unsupervised learning'' OR ''active learning'' OR ''deep learning'' AND ''attention model'')

The search results are displayed in Table 3. Even though this domain has been studied since 2000, the focus is put on the papers from 2011 to 2021 to depict current development in the field.

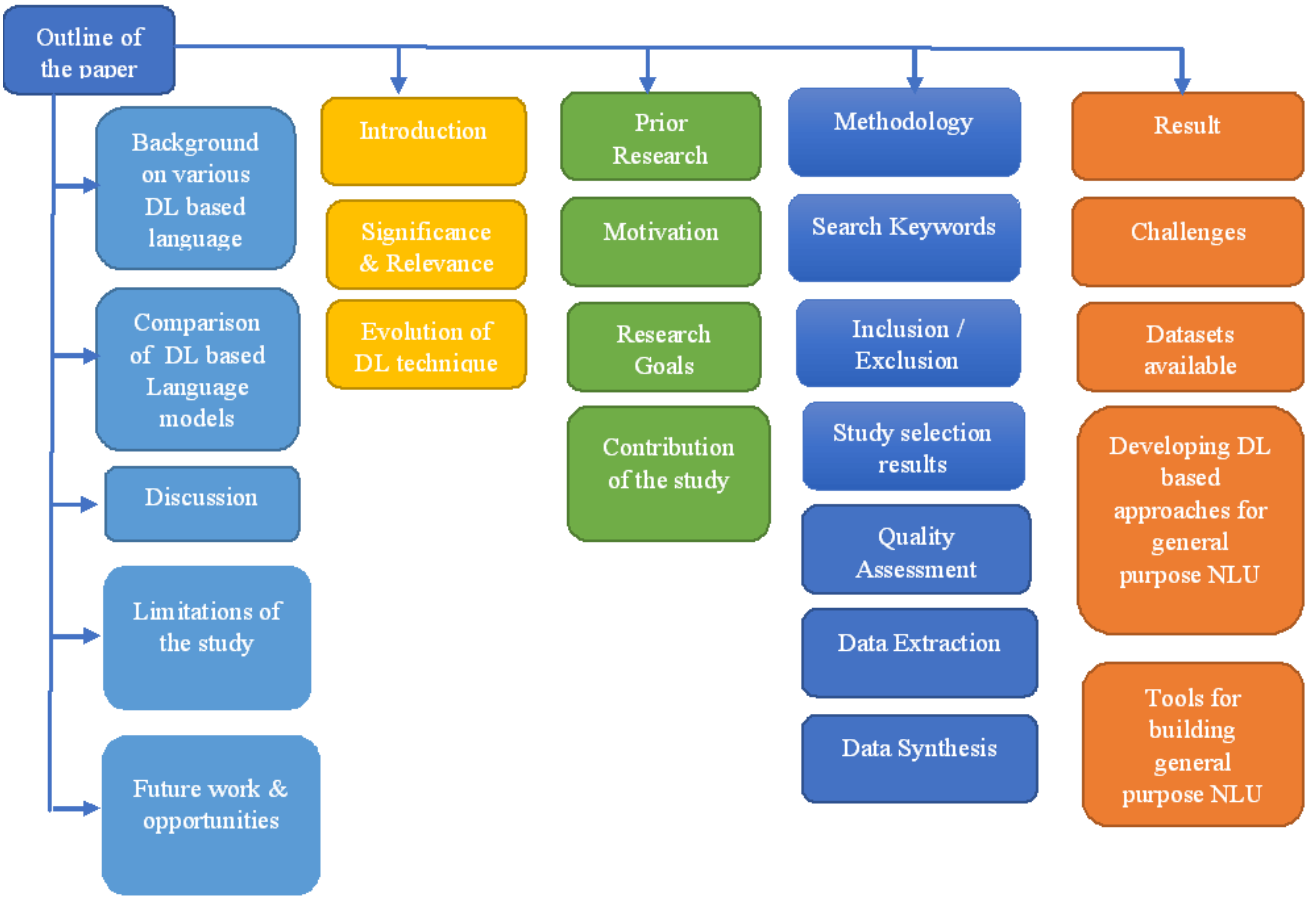


FIGURE 2. Outline of the paper.

TABLE 3. Literature databases search result.

Source database	Total Count	Relevant count after applying inclusion-exclusion criteria and duplicates removal
Scopus	197	48
Web of science	71	33
Emerald	72	02
IEEE Xplore	03	01
Google Scholar	438	22
Total	781	106

TABLE 4. Inclusion and exclusion criteria.

Criteria no.	Topic	Inclusion criteria	Exclusion criteria
1.	Language models	Focus on NLU tasks	Focus only on NLP tasks
2.	Multi-task learning	Focus on a combination of tasks	Focus on only a single task
3.	Year range	2011-2021	Before the year 2011
4.	Research questions relevance	Related to at least one research question	Not related to research questions

B. INCLUSION AND EXCLUSION CRITERIA

Research papers considered for this SLR must be relevant. The studies ranged from enhancing the techniques, building the frameworks for NLU, and catering to different application domains. The language selected is English, and it may be peer-reviewed. The inclusion and exclusion criteria to select the papers are as follows:

C. STUDY SELECTION RESULTS

The flowchart for choosing the pertinent papers for this SLR is shown in Figure 3. The search string was selected to earmark 781papers from the different databases mentioned in Table 4. After removing duplicates and applying inclusion and exclusion criteria remaining 106 research papers were considered for this SLR. Using snowballing techniques to include significant contributions, the total was increased to 115. Finally, after applying quality assessment criteria, 102 studies were selected for preparing a systematic literature review.

D. CRITERIA OF QUALITY ASSESSMENT FOR STUDY SELECTION

Quality assessment criteria ensure the relevance of research papers to riposte the RQs. The research studies were graded as 1 or 0 according to the criteria mentioned in Table 6.

TABLE 5. PIOC information (Population, Intervention, Outcome, Context).

Parameter	Meaning	Keywords
Population	Domains of applications	"Natural Language Understanding" OR "NLU"
Intervention	Software approaches/framework	"Active learning" OR "unsupervised learning."
Outcome	Element of significance	"multi-task NLU framework"
Context	Context of intervention	Attention model

TABLE 6. Quality assessment criteria.

Criteria	Score
Study provides results	1 if the criterion is TRUE else 0.
The study offers empirical proofs	1 if the criterion is TRUE else 0.
The study includes objectives and outcomes	1 if the criterion is TRUE else 0.
The study provides proper references	1 if the criterion is TRUE else 0.

The score 4 of quality is considered for conducting this SLR. Table 9 lists the quality assessment criteria.

E. DATA EXTRACTION

A concise overview of the data extraction process to answer the research questions from the studies is listed in Table 7.

F. DATA SYNTHESIS

Table 7 shows the data synthesis to address the Research Questions elaborately.

IV. BACKGROUND

Natural language understanding involves building language models, training these models, and testing them for accuracy. Various NLU tasks, such as question answering and NLI, can be cast as a classification problem. This section presents the TC tasks given in this study. Sentiment analysis is the method of extracting the polarity and lookout of customers' views. The problem can be expressed as a two-class or multiclass problem. A news classification system can aid consumers in opting into relevant news in the present by spotting incipient topics or making appropriate news suggestions depending on the reader's preferences. Topic classification is a job

TABLE 7. Categorization of the chosen Studies to answer research questions.

Category	Related data extracted from selected studies
Techniques used for combining text-related tasks in multi-task NLU	Challenges to deal with the trade-offs while combining tasks in multi-task NLU
Enhanced transformer encoding techniques	Fastformers, BERT, ALBERT, RoBERTa, OpenGTP, ToyBERT
Dataset and validation techniques	GLUE, superGLUE, Datasets, validation techniques in public domain
Combining approaches for MTL and transformer encoding	MT-DNN, Ernie, XLnet, UniLM
Benchmarking datasets and models developed by the researchers in the domain of NLU	Publicly available benchmarking datasets, models developed by the researchers in the domain of NLU.

that involves determining the overall theme or title of a document, whether a movie review is about viewer rating or revenue grossed over the specified period. Question-and-Answer (QA) Extractive and generative QA tasks are the types of QA tasks. Extractive Take extents in a document in SQUAD [19] as an example of a Text Classification task for a question and a collection of an appropriate answer. Each candidate's response is classified as correct or not correct by the algorithm. QA-NLI forecasts whether the meaning of one text can be predicted from the meaning of the other. A label belongs to entailment contradiction and is unbiased to a couple of Text units by an NLI system [20].

QA Extractive and generative QA tasks are the two types of QA tasks. Extractive Tasks are spread across the document length in SQUAD [23] as an example of a TC task given a question and a collection of probable responses. Each candidate's response is classified as correct or not correct by the algorithm. This study only discusses extractive QA, a text creation assignment that generates answers on the fly.

NLI forecasts whether the meaning of one text may be predicted from the meaning of the other. A text pair comparison is a generalized kind of NLI called paraphrasing. The problem of determining how likely one sentence is paraphrased from the other by comparing the semantic similarity of two sentences.

Neural Machine Translation - The objective of neural machine translation is to translate text by simulating the capabilities of the human brain. The goal is to translate a given source language into a target language retaining its meaning and intent. When translating, human brains first comprehend the sentence, then build a mental representation of the sentence, and finally convert this mental representation into a sentence in another language.

Through two modular processes of encoding and decoding, neural machine translation imitates the human translation process. The encoder turns source language utterances into semantic space vector representations. Depending on the semantic vectors produced by the encoder, the decoder constructs semantically identical phrases in the target language. RNN machine translation is a crucial basic model, and there

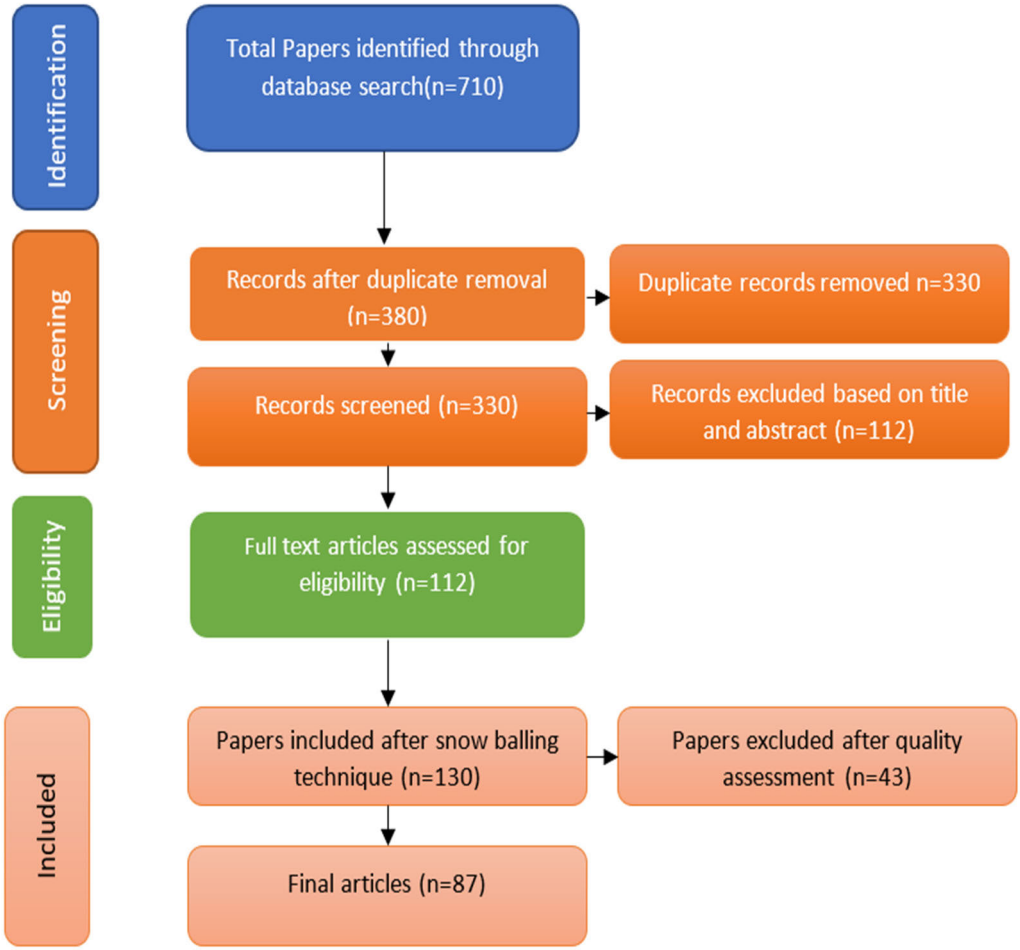


FIGURE 3. Flowchart for selection of relevant papers.

have been numerous advancements in advanced network topologies and unique model training methodologies.

Machine reading comprehension is the task of reading and comprehending the text by a computer program. The endeavour necessitates the creation of systems that can respond to queries after reading a document. This requirement has many uses, including allowing search engines to offer intelligent and correct responses to natural language inquiries by reading related text. Furthermore, reading comprehension machines can be used in virtual assistants to answer customer support questions after reading documents. It can also be utilized in the workplace to assist users in reading and processing emails or business papers, as well as summarizing pertinent information. Utilizing large-scale, manually annotated datasets has aided recent improvements in machine reading comprehension. This section looks at the numerous deep learning models that have been presented for text categorization problems. Based on their model structures, these models are put into the following categories: 1. Text is considered as a bag of words in feed-forward networks (section A) 2. RNN-based models are utilized to guess word dependencies by viewing the text as words in a specific order. (Section B) 3. For text categorization, CNN-based models are taught to identify text styles. (Section C) 4. Capsule networks deal with the problem of information loss in CNS pooling operations and have been used for text categorization (section D) 5. The attention mechanism is useful in constructing Dell models because it is effective in identifying correlated words in the text (section E) 6. Graph neural networks are designed to represent natural language's inherent graph structure. (Section F) 7. Hybrid models syndicate attention and texts to capture local and global features (section G) 8. Transformers are mainly employed for far more parallelization than RNNs, allowing for GPU-based training of huge language models (section H). These various types of models based on deep learning techniques is shown in Figure 4.

A. FEED-FORWARD NETWORKS BASED MODELS

Simple DL models for text representation include feed-forward networks. Despite this, they have a good level of accuracy on several TC benchmarks. Text is viewed as a collection of words in these models. These models acquire a vector representation for each word by word2vec [21]

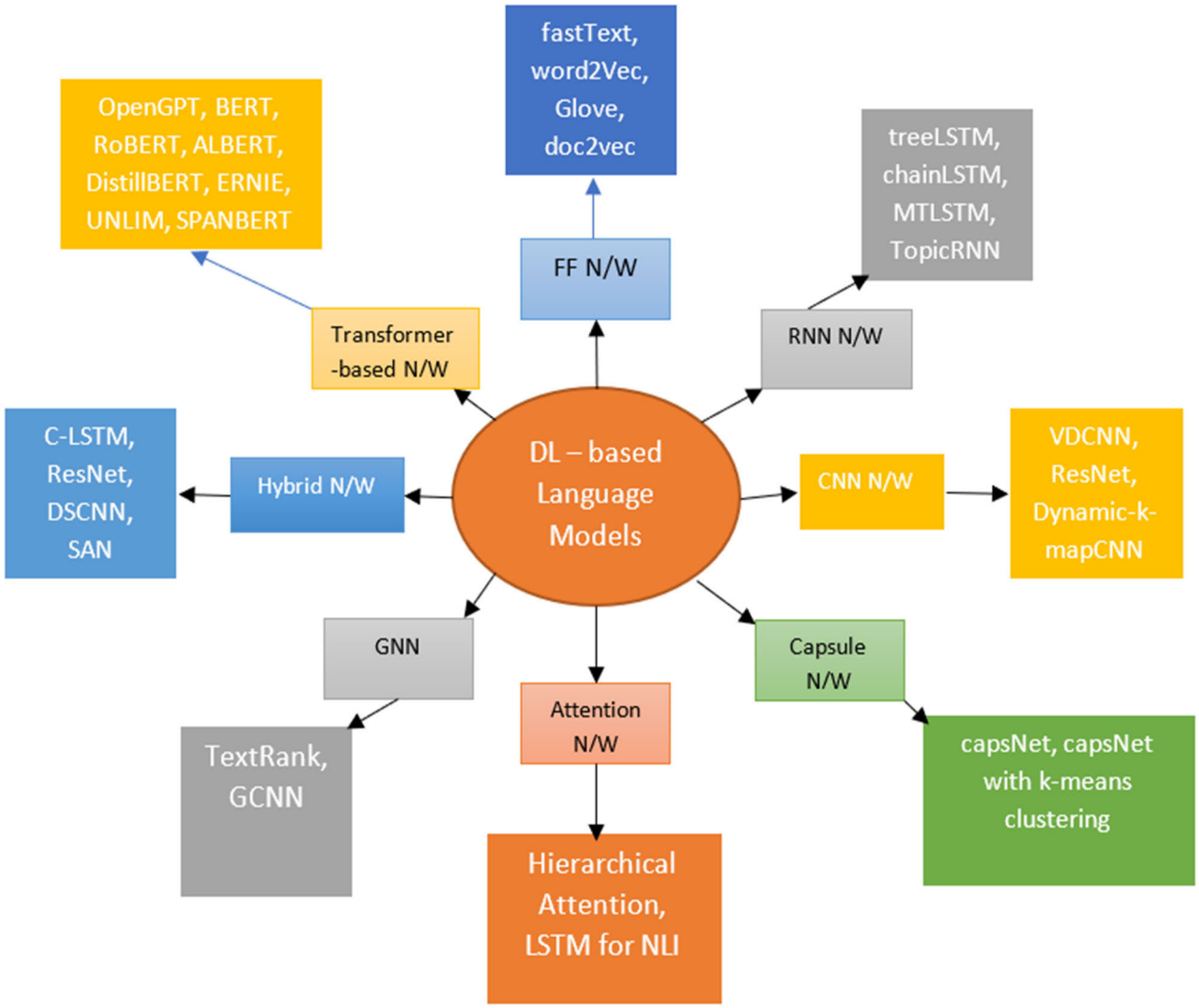


FIGURE 4. Types of language models based on Deep Learning.

or GloVe [22]. These are popular embeddings models. Joulin *et al.* [23] introduced another classifier called fastText. It is efficient and straightforward. A collection of n-grams is used as a supplementary feature in fastText to know the information about local word order. This technique proves efficient by reporting comparable results to the methods that utilize the order of words [24]. Le and Mikolove [25] introduced doc2vec, a method for learning fixed-length feature depictions of variable-length text, using an unsupervised algorithm.

B. RNN-BASED MODELS

Usually, the text is treated as an order of words in RNN-based models. The basic purpose of an RNN-based model for text categorization is to capture word relationships between sentences and text structure. Plain RNN-based models, on the other hand, do not perform as well as standard feed-forward neural networks. The Long Short-Term Memory (LSTM) model is one of many RNN variations meant to acquire long-term dependencies of words of sentences.

A memory cell with input, output, and forget gate is invented to remember the values over a stipulated time frame. LSTM models address the vanishing gradient and gradient exploding problems that plain RNNs suffer from using this memory cell. Tai and his colleagues, [26] to learn rich semantic representations, built a tree-LSTM model, which is a generalization of LSTM structured network topologies. Due to the syntactic structures of natural language, it combines words to form phrases, tree-LSTM is a more efficient model for NLP tasks than chain-LSTM. They demonstrate the performance of tree-LSTMs on two tasks: sentiment analysis and forecasting semantic relation between two sentences. Reference [27] Zhu *et al.* improved the performance of the chain-structured LSTM to its predecessors by storing many successor cells by a recursive process using memory cells.

By capturing useful information with various timeframes, the multi-time scale LSTM (MT-LSTM) neural network [28] is constructed to represent significant texts like sentences and documents. The hidden layers of a typical LSTM model are divided into numerous classes by MT-LSTM. At different times, each group is active and updated. MT- LSTM can successfully fit many documents in the model. MT- LSTM is reported to beat a collection of baselines, including LSTM and RNN-based text categorization models. RNNs are better at memorizing the local structure of word order, but they struggle with long-range type dependencies. For sentiment analysis, TopicRNN is said to outperform the RNN baseline. Other RNN-based models are also intriguing. Multi-task learning is used by Liu *et al.* [29] to train RNNs to utilize annotated data for training from various linked tasks.

C. MODELS BASED ON CNN

CNNs are taught to identify patterns in space, while RNNs are trained to detect patterns over time [30]. RNNs perform well in NLP tasks like RQA-POS tagging, which need an understanding of long-range semantics, but CNN performs well in situations where sensing local and location-independent patterns in the document is critical. These observed patterns could be significant sentences that convey a specific emotion. As a result, CNN has become the popular choice for common text categorization model designs. Kalchbrenner *et al.* [31] suggested one original text categorization algorithm. Based on CNN. The model is termed the Dynamic K- Max Pooling Model depending on the specified pooling technique. Dynamic CNN (DCNN) is a cable news network. The first stage of DCNN produces sentence metrics. The second stage involves a convolutional structure that integrates wide convolutional stages with dynamic convolutional settings.

The dynamic K-Max-pooling layers are utilized to construct a map of features over the entire sentence that captures various degrees of relatedness between the terms. The pooling parameter is selected at run-time based on the size of the sentence and convolution hierarchy level. For text categorization, Kim [32] suggests a considerably simpler CNN-based model. In this research, four distinct ways to learn word embeddings are compared:1. The model CNN-rand randomly initializes all word embeddings and then modifies them throughout training. 2. CNN-static, in which the pre-trained word2vec embeddings are utilized and remain static throughout model training, 3. In CNN-non-static, the word2vec embeddings are adjusted throughout training duration for individual task tasks. 4. In CNN-multi-channel, two sets of word embedding vectors, which are prepared using word2vec, and one model is modified throughout training, and the other remains unchanged. These models based on CNNs outperform previous models in sentiment scrutiny and question classification. Liu *et al.* [33] presented a novel CNN-based model that modifies the Kim- CNN architecture in two ways. A dynamic Max pooling approach collects additional detailed information from various document parts. Second, a hidden bottleneck layer is positioned in between the pooling and output layers to learn compact document representations to lessen the size of the model and improve accuracy. Rather than using low-dimension word vectors as input to CNNs, the researchers use high-dimension text to know the embeddings of short text areas of categorization in [33].

Prusa and Khoshgoftaar [34] offered a technique for encoding input text by using CNNs that significantly decreases the amount of memory used and the amount of training time necessary to acquire alphabet-level text data representations. In this method, the model grows with the number of characters, allowing better information from the text to be maintained in the augmented version.

There have been studies looking into how word embeddings and CNN architectures affect the model's performance. Conneue *et al.* [35] introduced a VDCNN model which uses deep architecture model for the task of text classification, which is inspired by resNets [36]. It uses modest convolutions and pooling operations and works directly at the character level. The performance of VDCCN improves as the depth is increased, according to this study. Deque *et al.* [37] changed the structure of VDCNN to match the limits of mobile platforms without sacrificing performance. They could reduce the model size by up to 20x with a 0.4 % to 1.3 % accuracy loss. Guo *et al.* [38] investigated the effects of word embeddings and recommended that weighted word embeddings be used in CNN model with multiple channels. Zhang and Wallace [39] studied various types of word embedding techniques and mechanisms, for pooling concluding that word2vec and GloVe perform better than one-hot vectors. Max-pooling is the best among the existing pooling approaches.

D. CAPSULE NEURAL NETWORK-BASED MODELS

CNN uses several layers of convolutions and pooling to classify pictures or text. Pooling operations detect significant features and minimize the computation complexity of convolution processes, but they miss spatial information and may misclassify items depending on their orientation or proportion. Hinton *et al.* [40] offered a new technique termed capsule networks to overcome the challenges of pooling (CapsNets). A capsule is a neuron collection whose activity vector shows many characteristics of a block or partial block. The length of the vector denotes the likelihood that the block exists, and the vector's orientation represents the block's properties. Capsules direct every capsule from the below layer to its best suitable parent capsule in the above layer, by available information in the network up to the last layer for categorization will stop directing. This task can be achieved using various algorithms like dynamic routing- by- agreement [41] or the Expectation-Maximization algorithm [42]. These networks have recently been used to classify text, with capsules customized by representing a line in the document or the entire document in the form of a vector. Kim *et al.* [43] developed a capsNet-based model with a comparable architecture. The model is made up of

four layers: 1. Documents are accepted by the input layer as a series of word embeddings, 2. Feature map is made by a convolutional layer and utilizes a gated-linear unit to retain certain information, 3. Local features are gathered by a convolutional capsule layer, 4. And finally, a text capsule layer that predicts class labels. Objects can be built more freely in text than in photographs, according to the authors. In contrast to the positions, the semantics of a document can remain unchanged, although the sequence of some lines of the document is changed. Ren and Lu [44] presented a new variation of capsNets that employs a combined architectural style among capsules and a novel k-means clustering-based routing algorithm. First, all codeword vectors in codebooks are used to create word embeddings. The lower-level capsules' features are then consolidated in high-level capsules utilizing k-means routing.

E. MODELS WITH MECHANISM OF ATTENTION

The way one pays attention to distinct sections of a photograph or related words of a single sentence motivates attention. Attention is becoming a central concept and tool in developing DL models for NLP [45]. It can be thought of as a vector of significant weights in a nutshell. To guess a word in a sentence, it is estimated how strongly it is related to the other words by using the attention vector, and by adding weighted values of the attention vector, the target value is predicted. This section examines some of the most widespread attention models that help build a new frame of mind. Yang *et al.* [46] suggested a network based on the mechanism of hierarchical attention for text categorisation. This model has two distinguishing features: one hierarchical structure that mimics the hierarchical structure of the document and two levels of attention mechanisms applied both at the word level and sentence levels, allowing it to pay differential attention to more and less significant parts while constructing the document representation. On six text categorization tests, this model surpasses earlier methods by a substantial margin. The hierarchical attention approach is extended to cross-lingual sentiment classification by Zhou *et al.* [47]. The document is modelled using an LSTM network in each language. Afterwards, the final sentiment analysis is performed using hierarchical attention in the sentence-level model. The attention models, designed at word-level, on the other hand, understand important words in each stage. Shen *et al.* [48] discovered a model which has self-attention mechanism for NLP, with directional and multi-dimensional attention between elements from input sequences. To learn sentence embedding, a high-weight neural net is utilized, which is exclusively dependent on the type of attention and has not consist of CNNs or RNNs. Liu *et al.* [49] provided an LSTM model for NLI that includes inner attention. To encode a text, this model employs a two-stage procedure. First stage sentence representation is generated using average pooling across word-level Bi-LSTM. After that, an attention mechanism is used to swap average pooling on the same phrase with superior representations. This technique pushed up the sentence representation at the first stage and further utilized attention to process the text.

F. MODELS BASED ON GRAPH NEURAL NETWORKS

Even though ordinary texts have a serial order, they also comprise inherent graph structures similar to parse trees that speculate the relationships based on syntax and semantics of the sentences. TextRank [50] is an original graph-based NLP model developed. The authors propose that a natural language text be represented as a graph G (V, E), with V denoting a collection of nodes and E denoting a set of edges between the nodes. Nodes characterize various text units that complete the sentences. Depending on the type of applications, edges can also be used to express multiple forms of relationships between nodes. Contemporary Graph Neural Network (GNN) is created by adapting DL methods. TextRank is one such method to graph data. Over the last few years, DNN models based on CNNs, RNNs, and autoencoders have been adapted to handle the complexity of graph data [51]. GCNs [52] and their derivatives are very prevalent among the numerous types of GNNs as they are operative and easy to mix with other networks, and they have reached optimal results in various applications. On graphs, GCNs are a more efficient variation of CNNs. To learn graph representations, GCNs pile layers of first-order spectrum filters trailed by an activation function, which is nonlinear. TC is a common application of GNNs in NLP. To infer document labels, GNNs use the interrelationships of documents or words [53]. For TC, Yao *et al.* [54] employed a GCNN model which uses CNN with graph networks. They learn a Text Graph Convolutional Network (Text GCN) for the corpus after creating a single text graph. It is based on words occurring together and word relations among the document. Text GCN starts learning with a one-hot representation of every word in the document and then further learns embeddings for both documents and documents. This method is supervised by known document class annotations. It's expensive to train GNNs for a colossal text corpus. Efforts have been made to reduce the cost of modelling by either dropping model complexity or adapting the model training techniques.

G. MODELS WITH HYBRID TECHNIQUES

Many hybrid models have been built to detect global and local documents by combining LSTM and CNN architectures. A Convolutional LSTM (C-LSTM) network is proposed by Zhou *et al.* [55]. C-LSTM uses a CNN to excerpt an arrangement of phrase (n-gram) representations, then input to an LSTM network to generate the sentence-level representation. For document modelling, Zhang *et al.* [56] suggest a Dependency Sensitive CNN (DSCNN). Chen *et al.* [57] used a CNN-RNN model to perform multi-label TC. A CNN is used by Tang *et al.* [58] to understand sentence representations that encode the inherent relationships between sentences. Xiao and Cho [59] considered a document in the specific order of characters rather than words and suggested

encoding documents using alphabet-based convolution with recurrent layers. When compared to word-level models, the proposed model achieved comparable results with many fewer parameters.

For learning word representations, Recurrent CNN [60] uses a recurrent structure for detecting long-range contextual dependence. To cut through the clutter, max-pooling is used to automatically select the most important terms for the text categorization task. In machine reading comprehension, Liu *et al.* [61] proposed a resilient Stochastic Answer Network (SAN) for reasoning using a multi-step approach. SAN incorporates multiple types of neural networks, such as networks with memory. Bi-LSTM, CNN, attention, and Transforms are all examples of transforms. The representations of the context for questions -answering system for the passages are attained using the Bi-LSTM component. Its attention method generates a passage representation that is question-aware. The passage is then stored in a working memory created by another LSTM. Finally, predictions are generated by an answer module that uses a Gated Recurrent Unit. Combining highway networks with RNNs and CNNs has been the subject of several studies. Information travels layer by layer in conventional multi-layer neural networks. With increasing depth, gradient-based DNN training becomes more complex. Highway networks [62] are designed to make deep neural network training easier. They permit the unrestricted flow of information across multiple levels over data highways, akin to ResNet [36] quick connections.

H. MODELS BASED ON TRANSFORMERS

The sequential processing of text is one of the computational obstacles that RNNs face. Even though RNNs are more sequential than CNNs, the computing cost of capturing associations among words in a phrase climb with the length of the sentence, much like CNNs. Transformers-based models [7] overcome this restriction by using self-attention to calculate an ''attention score'' for every word in a sentence of the document simultaneously, modelling the influence of each word on the others. Transformers, unlike CNNs and RNNs, allow for far more parallelization, allowing for the efficient training of huge models on massive volumes of data on GPUs.

Since 2018, several huge-scale Transformer-based Pre-trained LMs have emerged. Transformer-based models employ far deeper network architectures (viz., 48-layer based Transformers [63]). The pre-training of these models is also done on larger amounts of text to capture the context of the text representations.

Popular PLMs are classified by representation forms, model styles, pre-training tasks, and relevant tasks, according to the latest survey by Qiu *et al.* [64]. Autoregressive PLMs and autoencoding PLMs are the two types of PLMs. OpenGPT [8], a unidirectional model that forecasts a text verbatim from either left to right direction or vice-versa, with every word prediction based on earlier predictions, being one of the early autoregressive PLMs. By adding linear classifiers for relevant tasks and adjusting labels related to tasks, OpenGPT can be tailored to downstream tasks like TC. BERT [9] is one of the baselines autoencoding PLMs. Contrasting OpenGPT, which forecasts words based on past forecasts, BERT is normally trained by utilizing the masked language modelling (MLM) task, which arbitrarily masks some part in a text sequence and then improves them independently using encoding vectors produced by a Transformer, which processes text in both directions. RoBERTa [65] is a more robust version of BERT trained with more data. ALBERT [66] reduces BERT's memory use while increasing its training pace. DistillBERT [67] uses knowledge distillation technique throughout pre-training to reduce BERT's size by almost half while preserving its original capabilities and speeding up inference by a factor of two. SpanBERT [68] is a BERT extension that improves the representation and prediction of text spans. External knowledge bases are incorporated into ERNIE [69] to improve performance. XLNet [70] combines the concepts of autoregressive models such as OpenGPT and BERT. As previously stated, OpenGPT learns text representation for natural language creation using a left-to-right Transformer, whereas BERT employs a transformer, which processes text in both directions, for natural language interpretation. Unified Language Model (UniLM) [71] is a model for understanding and creating natural language. UniLM has been pre-trained on different types of language modelling tasks which are not related to the direction of parsing. XLNet uses a transformation operation throughout the pre-training phase to allow words from both the left and right sides of the context to be included, making it a bi-directional autoregressive model. In the Transformers model, the transformations are performed by utilizing a specific attention mask. To facilitate position-aware word prediction, XLNet provides a two-stage self-attention schema. This schema is based on how distributions of words change dramatically, relevant to word location.

As previously stated, OpenGPT learns text representation for natural language creation using a left-to-right Transformer, whereas BERT employs a transformer that can process text in both directions for natural language interpretation. The Unified Language Model (UniLM) [71] is a model for understanding and creating natural language. UniLM has been pre-trained on different language modelling tasks which are not direction specific. A common Transformer architecture fused with the specific self-attention masks are used to create the unified modelling. UniLM [71] has reached new SOTA performance for natural language interpretation and generating tasks, outperforming prior PLMs. The performance analysis of significant language models, along with the techniques employed, is summarized in Table 8.

V. RESULTS

This section provides the answers to the research questions. Major issues and challenges are outlined in multitasked-based deep learning language models.

TABLE 8. Performance analysis of significant language models.

Reference	Dataset	Techniques	Application	Performance (Accuracy)
[9]	AGNews	BERT	Text classification	92.7
[23]	AGNews	fastText	Text classification	92.5
[26]	SNLI	treeLSTM	Inference	85.7
[27]	Amazon-2, Yelp-2	chainLSTM	Product review, Sentiment analysis	94.39, 95.12
[35]	SST-2	VDCNN	Sentiment analysis	84.5
[40]	AGNews, SogueNews	CasuleNetB	Text classification	92.39, 97.25
[51]	20News	GNN	Text classification	88.5
[61]	SNLI	SAN	Inference	85.6
[65]	SNLI, SST-2, SQuAD 2.0	RoBERTa	Inference, Sentiment analysis, QA system	92.6,95.6, 82.3
[66]	SST-2	ALBERT	Sentiment analysis	95.2
[68]	SQuAD 1.1, SQuAD 2.0	SPANBERT	QA System	94.6, 73.7
[70]	AGNews	XLNet	Text classification	95.5
[71]	AGNews, IMDB	ULMFiT	Movie reviews	82.1, 95.4

RQ1. What are the different algorithmic approaches available in combining MTL to improve the accuracy of the proposed framework for NLU?

The major challenge in NLU research involves language representation of various tasks for training. The previous method, rule-based NLU, relied on human-generated characteristics. As a result, it cannot recognize unexpected words and requires a significant amount of time for features engineering. NLU, based on deep learning, on the other hand, extract features and learns language representations automatically. MTL models, which were created from the base of single-task learning, which learns a generalized representation of while preventing overfitting of a given individual task, have gained a lot of attention in recent years.

- MT-DNN [72]- This model and MTL models based on the bidirectional transformer are two examples of MTL models. MT-DNN learns many tasks at the same time. Instead of travelling through both common and task-specific levels, this approach, on the other hand, uses a method of knowledge distillation that teaches a smaller multi-task model using larger single-task models. MTL models, on the other hand, have several difficulties when it comes to fulfilling the essential duty of maintaining overall accuracy. The accuracy of the model can be affected due to joint training with the equal weights.

  Furthermore, during MTL, the model is unable to master the main job fully. These drawbacks are compensated by presenting the SIWLM (Sequential and Intensive Weighted Language Modeling) scheme, inspired by Yang *et al.* [73]. SIWLM comprises two types of learning: sequential weighted learning (SWL) and intensive weighted learning (IWL). Core and supplementary tasks in SIWLM have an initial task weight, and all tasks are independently changed during training. The loss functions are multiplied by these altered weights. The ideal weight for each job in the GLUE datasets can be calculated. The MTDNN-SIWLM attains equivalent or better performance on all GLUE datasets when that weight is applied.
- **Task Interference**- Another challenge is tackling task interference in multi-task learning. Task interference can be seen as a paradox of invariance vs sensitivity: essential information for one task may be futile information for the other, resulting in possibly conflicting aims while training multi-task networks leading to poor overall performance.
- **Attention Strategies** - The studies [76-77] propose two separate task attention strategies. To begin, task-specific data-dependent modulation signals boost or decrease neural activity. Second, task-specific Residual Adapter extracts task-specific information blended with the symbols created by a common task structure. This method enables us to learn a common symbol system that serves all activities while also collaborating with a task-related processing to develop more complex task-related structures. Different task attention strategies may lead to minimum task interference.
- **Common features learning- [76]** describes a method for learning a low-dimensional representation used in various applications. The method uses a new regularizer to manage the number of learning features common to all tasks, which builds on the well-known 1-norm regularization problem. The authors show that this problem is comparable to a convex optimization problem and propose an iterative solution. The approach has a straightforward interpretation: it alternates between supervised and unsupervised steps, in which the latter learns representations that are common across tasks, and the former learns task-specific functions using these

representations. There is scope to extend this approach and develop new iterative algorithms.

While deep learning and deep reinforcement learning (RL) has shown promise in enabling computers to perform complicated tasks, existing approaches data needs make it challenging to acquire a wide range of capabilities, especially when each work is learned independently from scratch. To solve multi-task learning difficulties, a natural strategy is to train a network on multiple tasks simultaneously to uncover shared structure across the tasks in a way that is more efficient and effective than tackling tasks individually. [77] It may obtain the hypothesized benefits of multi-task learning without the cost of ultimate performance if optimization issues are properly addressed for multi-task learning.

All the NLP tasks, typically QA, content summarization, NLI, formulate the candidate tasks are considered a benchmark for the Natural Language Decathlon (decaNLP). [72]. This study proposes an approach to recast all jobs as a series of questions to be answered in a certain context. A novel multi-tasks question answering network (MQAN) is proposed, which learns decaNLP tasks without fine-tuning any network [72]. The knowledge representation other than the QA pair needs to be evaluated for MTL- NLU.

The approach suggested in MT-DNN [72] appears to be practical due to its exhaustive nature. The accuracy of the model is sensitive to strategies adopted for effectively handling task interference. Higher amount of task interference can adversely affect the performance of the model.

RQ2 Which are the standard benchmarking datasets for MTL- NLU tasks evaluation?

GLUE [78] and its successor SuperGLUE [79] are the standard benchmarks for evaluating a model's performance on a set of tasks rather than a single task to keep a broad picture of NLU performance. GLUE comprises of the following ingredients:

1. A benchmark dataset consists of nine text-related NLU tasks based on specified datasets and chosen to deal with a wide variety of dataset parameters like size, type and hardness of tasks.
2. A analytical dataset for evaluating and analyzing the framework accuracy in natural language concerning a wide range of linguistic phenomena and a public leader board for tracking performance.

   In recent years, new pre-training and transfer learning models and approaches have resulted in significant performance gains across various language comprehension tasks. The GLUE benchmark, introduced in 2019, provided a single-number metric summarising progress on various such tasks.

   However, performance on the benchmark has lately approached that of non-expert humans, indicating that there is limited potential for further development.

   In [79], SuperGLUE, a new benchmark inspired by GLUE that includes a new set of more difficult language comprehension problems, enhanced resources, and a new public leaderboard, is established as a de-facto benchmarking standard for MTL-NLU. Table 9 enlists the details about these datasets.

   The field of benchmarking datasets for natural language understanding is rapidly evolving. There is definitely a future possibility to include more tasks related to NLU.

TABLE 9. Detailed information about GLUE and superGLUE SOURCE [80,81].

Dataset	Corpus	Task	Metric	Domain
GLUE	CoLA	Single sentence	Matthews coefficient	Mixed
	SST-2	Sentiment Analysis	Accuracy,	Reviews
	MRPC	Similarity and Rewording Tasks	F1 Score, Accuracy	News items
	STS-B	sentence similarity	Pearson or Spearman coefficient	Mixed
	QQP	Rewording	F1 score, Accuracy	Question-Answer
	MNLI	Inference	Accuracy	Mixed
	QNLI	Inference	Accuracy	Wikipedia
	RTE	Inference	Accuracy	Wikipedia, News items
	WNLI	Inference	F1 Score	Novels
superGLUE	BoolQ	Question-Answer	Accuracy	Google queries, Wikipedia
	CB	NLI	Accuracy, F1	Mixed
	COPA	Question-Answer	Accuracy	blogs, photography encyclopedia
	MultiRC	Question-Answer	F1	Mixed
	ReCoRD	Question-Answer	F1/EM	news (CNN, Daily Mail)
	RTE	NLI	Accuracy	News, Wikipedia
	WiC	WSD	Accuracy	VerbNet, Wikipedia
	WSC	coreference	Accuracy	Novels

RQ3. Which are the learning methods for improving the learning performance of NLU in combining multiple tasks?

Text-based games use natural language to recreate worlds and interact with players. Recent research has used them as a testbed for autonomous language-understanding bots, with the premise that comprehending the meanings of words or semantics is critical to how humans understand, reason, and act in these worlds. However, it's unclear how much semantic understanding of the text is used by artificial agents.

- **Semantics**-The studies [82-83] conducted in the context of ZORK-I, a text adventure game, revealed contradictory evidence to this basic principle and proposed an improvement in the structure of the NLU unit. It probes the extent of semantics in the reinforcement learning agents used in text adventure games. It also demonstrates

the techniques to adjust the semantics plugged into the system.

- **In joint training** in the study [82], the authors proposed a deep RNN with partial relative dialogue memory by mutually training the NLU unit and System Action Prediction unit. This approach is different from the conventional one where the NLU unit and SAP unit are trained in a pipeline. The major drawback of noisy NLU badly affecting the performance of SAP unit is successfully handled.
- **Gated-Attention**-The study [83] proposed Gated-Attention Reader- an integrated model with a new attention mechanism based on multiple interaction. This technique permits the user to create a query-related representation of tokens for accurate response prediction. This study improves on attention mechanism to improve the accuracy of the results. The study [84] proposed UMLFiT, a technique to fine-tune any language model. It is a transfer learning method applicable to any task. This method effectively outperforms the existing models on various TC tasks, and the prediction errors are significantly reduced by 19 to 24 %. It also requires only 100 labelled samples to match the training from scratch on 100x more samples. In the study [85], a novel training technique is used to train CNN in news comprehension experiments that use news articles and summarized bullet points in the form of QA pairs for training. This method adds a new training method for QA systems.
- **Frame semantics**-The study [86] proposed a new theory—Frame semantics for the English language. It addresses the issue of frame-semantic analyzing by means of a statistical method which treats lexical targets in their sentential context and predicts frame semantic structures is handled. This model is based on latent variables and semi-supervised learning to remove disambiguates from frames.
- **Construction Grammar**-In the study [87], Construction Grammar (CG) is used along with AI to represent the knowledge for a deep understanding of a text. The experiments involve Winograd Schema (WS) – a major test for AI. The results showed that the proposed CG approach has more potential for task resolution in the deep understanding of natural languages.
- **Transformer-based representation**-The study [88] summarizes the latest transformer-based models related to NLP models. It provides a thorough explanation and working of various designs, relative assessments, and forthcoming guidelines in NLP.
- **GROW Model**-In the study [89], a method is proposed with a pro-active attitude driving the dialogue to reach different coaching goals for elderly users. It used the communication supporting technique based on GROW model.
- **FastText**-The study [90] evaluates the word embeddings techniques for the Nursing care domain. It proposes an automatic labelling framework for dialogues between patients and nurses to record care activities. It also pointed out fastText as a better word embeddings model by achieving 0.79 as an F1 score measure.
- **NLU Services** -Study [91] describe the functioning of NLU services and their role in the general architecture of a chatbot. It presents the comparison of existing NLU services like DialogFlow (Google), Watson (IBM), and Alexa (Amazon). Study [92] focuses on the Italian language as a non-English language for using the NLU engine.
- **RASA-NLU**-In the study [93], a functional framework is proposed, and the principles of RASA NLU are introduced. It combines neural nets (NN) and RASA NLU for implementing the systems based on entity extraction after recognizing intents. The findings state that RASA NLU outperforms NN for a single word, but NN has better integrity for segregated words classification. The study [94] highlights an approach to syndicate various types of semantics and language models while pre-training and fine-tuning stage for the improvement in the accuracy of prediction.
- **OOD Input**- The study [82] deals with the problem of Out of Domain (OOD) input. OOD input may lead to system failure. A novel method to generate high-quality pseudo-ODD samples is proposed. The training is done using a generative adversarial network. An auxiliary classifier is used to regularize the generated OOD samples. The results show improvement in OOD detection and efficient utilization of unlabeled data.
- **Semantic Vector** - In the study [75], semantic vector learning for NLU is explored. An NLU embeddings model is focused in the light of the understanding relationship between unstructured text and corresponding structured semantic knowledge. The contribution method is to create matching vector with relevant semantic frame. The applications of this model include visualizing distance-based semantic search and similarity-based intent classification and re-ranking.
- **Selective classifier method**-The study [95] tries to solve the problem of domain-specific training. A new technique called a top-down particular multi-classifier system ensemble model is proposed, and it offers a significant improvement over the word embeddings method. The study [96] demonstrates an effective user interface design for NLU based stock analysis system. The system is based on RNN with hyperparameter tuning. The study [97] substantiates the theoretical analysis of NLU as a methodological problem. The study [98] underlines the unsuitability of NLP approaches for NLU systems. The conventional methods based on statistical-based supervised learning must be replaced by the holistic cognitive modelling approach. A new paradigm called Onto-Agent- based on human cognition is proposed.

- **Learning without forgetting** - In the study [99], the problem of adding new tasks with existing training data is tackled. The proposed method – learning without forgetting with CNN is capable of learning new tasks with new task data; while preserving the original capabilities. This method is better than feature extraction and fine-tuning adaptation in multi-task learning.
- **Garden Path sentences**- The study [100] addresses the problem of Garden Path (GP) sentences. It demonstrates the effective use of Popescu's model of NLU systems that states that syntactic, semantic, and cognitive background knowledge is essential in training data.
- **Dynamic Integration of background knowledge**-The study [101] introduces a new architecture for the active integration of explicit background knowledge in the NLU model. Experiments on Document Question Answering (DQA) and recognizing textual entailment (RTC) demonstrate the effectiveness and flexibility of the proposal.
- **Representation conversion**-The study [102] describes and evaluates various methods to the alteration of the gold standard corpus data from Stanford-typed Dependencies and Penn-style constituent trees to the modern English Universal dependencies (UD 2.2). The outcomes show a reduction of 1.5% errors.

Among these approaches, the method of gated attention is prevalent in the current literature due to its efficiency.

RQ4. Which techniques are effective for reducing the need for huge annotated data samples?

- **VIRAAL**- The study [105] proposes a new approach – Virtual Adversarial Active Learning (VIRAAL) to reduce annotation efforts in NLU systems. It uses a semi-supervised model that regularizes the model through local distribution smoothness. The importance of Entropy-based active learning is underlined by querying more informative samples without requiring additional components. Results show that this method is robust on multi-task NLU training.
- **Sub-modularity-inspired data ranking function** -The study [106] tries to address the problems for small domains, which requires a huge amount of domain-related training data. The data selection technique is proposed in a low-data regime that enables training with fewer labelled data.
- **HUMOD**- In the study [107], the unavailability of a common metric to evaluate the replies against human judgment is handled. This study contributes by developing a benchmark dataset with human annotation and diverse responses. HUMOD- a high-quality human-annotated movie dialogues dataset. It is created from the Cornell movie dataset. Detailed analysis on the structure of dialogues and human perception score in comparison with existing models is presented.
- **SSG Framework**- The study [108] proposed a dual Learning technique for semi-supervised NLU. This study introduces a dual-task, semantic to sentence generation (SSG) framework. It enables the NLU model to fully use labelled and unlabeled data. The framework achieves impressive results on publicly available datasets like ATIS and SNP.
- **Auxiliary Tasks** In the study [109], a novel approach of utilizing auxiliary tasks to provide additional supervision of the main task to compensate for the data paucity is proposed. It addresses the issue of assigning and optimizing importance weights to auxiliary tasks. A weighted likelihood function of auxiliary task is formed as a surrogate before the main task leading to the reduced need for additional training data.

The VIRAAL [72] method is more effective in reducing the need for huge annotation samples for training.

RQ5. Which are prevalent knowledge representation techniques for MTL- NLU?

- **Word embeddings** - Zhang and Wallace [39] studied the effects of several word-embedding methods and pooling mechanisms, concluding that word2vec and GloVe are better than one-hot vectors and that Max-polling is the best among the existing pooling approaches.
- **ELMO** - The Elmo [6] model performs substantially better than the word2vec approach because it captures contextual information in knowledge representation.
- **EntityNLM**-The study [110] proposes a knowledge representation method for tracking and evolution of the introduction of entities present in lengthy documents. The technique is called EntityNLM. It can model dynamic entities, dynamically update representations and contextually generate mentions of the entities. Different tasks like language modelling, conference resolution, and entity prediction outperform the baseline.
- **Model Pruning**-The study [111] proposes a technique to compress bulky language models while preserving information for a explicit task. A specific layer is selected among various layers to prune the language model.
- The model pruning [111] technique is more efficient and widely used method in compacting knowledge representation.

VI. DISCUSSIONS

This study examined a significant number of research publications (116 studies, to be specific) on various aspects for building language models created using deep learning approaches in this SLR. Several critical questions regarding Language models were discussed in this review. The summarized points are as follows -

- FFNN (Feed-forward neural networks) see input text as a bag of words. The RNNs can understand word ordering. The CNNs are strong in recognising patterns like significant words. The attention mechanisms are great at identifying associated words in the input text, Siamese Neural Nets s are utilised for text similarity

tasks and, GNNs performs better if natural language structures are beneficial to the intended task. But the transformer-based models outperform all other models.

- The architecture of the framework is determined by the intended task, ready existence of the annotated samples and the restrictions of application domain. These requirements can only be fulfilled by hybrid architecture language models.
- Existing studies explore MTL-based language models, which use a supervised learning paradigm. Very few studies combine other techniques with MTL to focus on unsupervised learning. There is scope to combine transformer-based representations with MTL for improving the efficiency and accuracy of the language framework model.
- There is scope in MTL to efficiently handle the issue of out-of-domain task detection and task interference by using the active learning technique. The availability of annotated samples is less in NLU tasks. There is scope for developing approaches to reduce the need for more annotated examples using active learning techniques.
- Pretrained Language models (PLM) improved performance for all text related tasks and autoencoding PLMs are more commonly used language models due to its efficiency.
- The layers related to the particular task can be trained independently or jointly with the PLM, depending on the existence of domain related labels. Multi-task fine-tuning is an excellent alternative for leveraging labelled data from related domains.
- The size of existing transformer-based models is humongous. There are many approaches available for the compaction of transformer-based representation. The optimal strategy for text encoding for NLU is not established in the current literature.
- Fusing commonsense knowledge into DL models has the probability to boost model efficiency dramatically, similar to how humans use commonsense understanding to complete various jobs. A generic QA system, for example, could answer queries regarding the real world. When there isn't enough information to address a problem, commonsense knowledge can aid. Machine learning-based systems can reason about unknowns using "default" conventions, similar to how people do, by using commonsense.

While DL models have shown potential on challenging benchmarks, the functioning of most of these models is not interpretable. For instance, it is still unclear how few models outperform others on one dataset while underperforming on others? What things exactly have deep learning models discovered? Although the attention mechanisms offer some perception into these concerns, a comprehensive analysis of these models' underlying behaviour and dynamics is still absent. Greater knowledge of these models' theoretical elements can aid in the development of better models tailored to diverse text analysis settings.

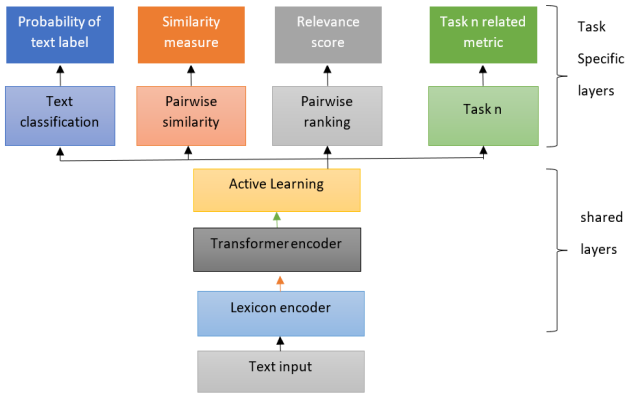


FIGURE 5. Proposed framework for multi-task NLU.

A. PROPOSED FRAMEWORK

This study started the SLR to answer five Research Questions to lay the groundwork for future study in the field of NLU. These RQs served as the cornerstone for the planned multi-task NLU framework. Figure 5 depicts the framework proposed.

As depicted in Figure 5, the proposed framework is derived from BERT and has three bottom layers, which are common for all the tasks, and the top-end layers show task-related representation and output. The input is text converted as word embeddings. The transformer encoder then uses self-attention to get contextual information for each word and creates a string of contextual embeddings in the next stage. The exchanged semantic presentation is used to achieve required goals in multi-tasks. The active learning technique is further used to obtain more information from fewer but more informative examples, reducing the demand for more samples.

The training of the proposed framework includes stages like pre-training and multi-task learning. The pre-training stage is similar to that of the BERT model. Masked language modelling and next sentence prediction are used to learn the parameters of both the encoders. The model's parameters are learned using the stochastic gradient descent (SGD) method in the final stage.

This framework is expected to perform well on the premise of combining multiple techniques.

VII. LIMITATIONS OF THE STUDY

The present DL language models employed in text classification tasks were examined and critically analyzed by this SLR. It offers comparisons and challenges for developing DL models for multi-task NLU. However, due to the scarcity of literature research and work in this field, as well as the wide range of DL models, finding and selecting relevant literature is a laborious, hard, and difficult job. To meet the required inclusion and exclusion criteria, the keywords used to search for valuable publications and procedures may vary or change.

One of the major restrictions of SLR for the domain is that, even though a systematic review process is followed, it cannot guarantee that all pertinent works of the domain were extracted. The most relevant electronic databases in computer science were included in the search databases. Another limitation is the authors' preconceived notions regarding the multi-task NLU procedure as a whole.

The suggested framework is in the design stages, and its empirical validation is beyond the scope of this SLR, but it demonstrates the long-term research goals.

VIII. FUTURE WORK AND CHALLENGES

Several concerns will need to be addressed in future investigations.

- **Outlier tasks**- To begin with, outlier tasks that are unconnected to other tasks are known to impede the performance of all activities when they are learned together. There are a few strategies for reducing the harmful consequences of outlier tasks. However, there are no defined approaches or theoretical assessments to investigate the detrimental consequences. This issue is a critical issue that requires more research to make MTL safe for human use.
- **Learning methods** - Deep learning has emerged as a prominent strategy in various fields, with various MTL deep learning models presented in the feature alteration, low-rank, task bunching, and task relatedness learning methods. As previously said, the majority of them simply have hidden common layers. This method is effective when all the tasks are linked, but it is prone to outlier tasks, significantly degrade performance.
- **Security** - The resilience of this multi-task DNN framework is to be checked against various types of attacks.
- **Extension**- Finally, most studies to date have focused on supervised learning tasks, with only a few focusing on learning types like unsupervised, semi-supervised, active, and reinforcement to those non-supervised learning problems; it is logical to modify or extend various multi-task methodologies. It is believed that such adaption and extension will necessitate greater effort in developing relevant models.

IX. CONCLUSION

The majority of the issues that MTL faces today are the same challenges it has encountered for the past two decades. The SLR aims to investigate contemporary DL-based language models for multi-tasking NLU to find areas where progress can be made. While it is still usually accepted that task relatedness leads to good bias, Caruana [112] demonstrated that certain inductive biases can be harmful, and while there is no strong universal notion of measuring it. The underlying difficulty of task interference, in which MTL is hampered by a plethora of complex and competing goals. Deeper and more general strategies for task selection and assessment are still needed. As more study into the consequences of MTL is conducted, it is critical to continue to improve the understanding of task connection and selection.

The SLR examines and analyses DL-based language models for multi-tasking NLU research by:

- Identifying the problems with existing language models for multi-tasking NLU
- Recognizing the necessity to combine supervised and unsupervised learning paradigms
- Exploring the need to combine transformer-based representation with word embeddings presentation for the text.
- Identifying model compression strategies that can be used to lower the size of the dataset.
- This area of study for various application areas, including conversational AI, chatbot-based systems for education, legal, stock market, and customer service, to name a few, is getting investigated.

The findings suggest that the hybrid model, which combines diverse strategies such as MTL and active learning, is given more consideration because of its effectiveness in managing text-related tasks for NLU. It is also noticed that the merging of supervised and unsupervised paradigms receives less attention. As a result, building a multi-task NLU combination model is a promising prospect. There's still scope to figure out how to make MTL models that are both resilient and capable of enabling the next generation of general AI.

REFERENCES

[1] Kommandotech. (2021). *Astounding Artificial Intelligence Statistics for 2020*. Accessed: Oct. 22, 2021. [Online]. Available: https://kommandotech.com/statistics/artificial-intelligence-statistics/

[2] Botcore. (2018). *Chatbots: The Past, Present, and Future*. Accessed: Oct. 22, 2021. [Online]. Available: https://botcore.ai/blog/chatbots-the-past-present-and-future

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Nov. 1985.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[6] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.

[7] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[8] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. *Improving Language Understanding by Generative Pre-Training*. Accessed: Oct. 22, 2021. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, and P. Dhariwal, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[11] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "GShard: Scaling giant models with conditional computation and automatic sharding," 2020, *arXiv:2006.16668*.

[12] Developer. (2021). *Announcing Megatron or Training Trillion Parameter Models Rivavailability*. Accessed: Oct. 22, 2021. [Online]. Available: https://developer.nvidia.com/blog/announcing-megatron-for-training-trillion-parameter-models-riva-availability/

[13] G. Marcus, "The next decade in AI: Four steps towards robust artificial intelligence," 2020, *arXiv:2002.06177*.

[14] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A new benchmark for natural language understanding," 2019, *arXiv:1910.14599*.

[15] E. S. Dos Reis, C. A. Da Costa, D. E. Da Silveira, and R. S. Bavaresco, "Transformer aftermath review," *Commun. ACM*, vol. 64, no. 4, pp. 154–163, Apr. 2021.

[16] Y. Zhang and Q. Yang, *An Overview of Multi-Task Learning*, vol. 5. London, U.K.: Oxford Univ. Press, 2018, pp. 30–43.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. N. Gomez, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5999–6009.

[18] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in SE," *Guidel. Perform. Syst. Lit. Rev.*, vol. 3, pp. 1–44, Oct. 2007.

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," 2016, *arXiv:1606.05250*.

[20] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 1–8.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[22] J. Pennington, R. Socher, and C. Manning, "GloVE: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, vol. 2014, pp. 1532–1543.

[23] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "FastText.Zip: Compressing text classification models," 2016, *arXiv:1612.03651*.

[24] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2. 2012, pp. 90–94.

[25] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[26] K. Sheng Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," 2015, *arXiv:1503.00075*.

[27] X. Zhu, P. Sobihani, and H. Guo, "Long short-term memory over recursive structures," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1604–1612.

[28] P. Liu, X. Qiu, X. Chen, S. Wu, and X.-J. Huang, "Multi-timescale long short-term memory neural network for modelling sentences and documents," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2326–2335.

[29] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," 2016, *arXiv:1605.05101*.

[30] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[31] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 1–11.

[32] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1–62.

[33] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 115–124.

[34] J. D. Prusa and T. M. Khoshgoftaar, "Designing a better data representation for deep neural networks and text classification," in *Proc. IEEE 17th Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2016, pp. 411–416.

[35] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," 2016, *arXiv:1606.01781*.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[37] A. B. Duque, L. L. J. Santos, D. Macédo, and C. Zanchettin, "Squeezed very deep convolutional neural networks for text classification," in *Artificial Neural Networks and Machine Learning* (Lecture Notes in Computer Science). Munich, Germany: European Neural Network Society, 2019.

[38] B. Guo, C. Zhang, J. Liu, and X. Ma, "Improving text classification with weighted word embeddings via a multi-channel TextCNN model," *Neurocomputing*, vol. 363, pp. 366–374, Oct. 2019.

[39] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015, *arXiv:1510.03820*.

[40] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proc. Int. Conf. Artif. Neural Netw.* Espoo, Finland: Springer, 2011, pp. 44–51.

[41] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.

[42] S. Sabour, N. Frosst, and G. Hinton, "Matrix capsules with em routing," in *Proc. 6th Int. Conf. Learn. Represent.*, 2018, pp. 1–15.

[43] J. Kim, S. Jang, and S. Choi, "Text classification using capsules," *Neurocomputing*, vol. 376, no. 1, pp. 214–221, Feb. 2018.

[44] H. Ren and H. Lu, "Compositional coding capsule network with K-means routing for text classification," 2018, *arXiv:1810.09177*.

[45] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[46] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *Proc. NAACL*, 2016, pp. 1480–1489.

[47] X. Zhou, X. Wan, and J. Xiao, "Attention-based LSTM network for cross-lingual sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 247–256.

[48] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "DiSAN: Directional self-attention network for RNN/CNN-free language understanding," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5446–5455.

[49] Y. Liu, C. Sun, L. Lin, and X. Wang, "Learning natural language inference using bidirectional LSTM model and inner-attention," 2016, *arXiv:1605.09090*.

[50] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 404–411.

[51] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," 2019, *arXiv:1901.00596*.

[52] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[53] G. Cucurull, A. Casanova, A. Romero, P. Liü, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[54] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7370–7377.

[55] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A C-LSTM neural network for text classification," 2015, *arXiv:1511.08630*.

[56] R. Zhang, H. Lee, and D. R. Radev, "Dependency sensitive convolutional neural networks for modeling sentences and documents," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1–11.

[57] G. Chen, D. Ye, E. Cambria, J. Chen, and Z. Xing, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," in *Proc. IJCNN*, 2017, pp. 2377–2383.

[58] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1422–1432.

[59] Y. Xiao and K. Cho, "Efficient character-level document classification by combining convolution and recurrent layers," 2016, *arXiv:1602.00367*.

[60] R. Wang, Z. Li, J. Cao, T. Chen, and L. Wang, "Convolutional recurrent neural networks for text classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–6.

[61] X. Liu, Y. Shen, K. Duh, and J. Gao, "Stochastic answer networks for machine reading comprehension," 2017, *arXiv:1712.03556*.

[62] R. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.

[63] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multi-task learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

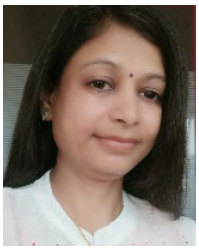
[64] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," 2020, *arXiv:2003.08271*.

[65] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[66] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.

[67] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[68] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," 2019, *arXiv:1907.10529*.
[69] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, "ERNIE: Enhanced representation through knowledge integration," 2019, *arXiv:1904.09223*.
[70] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pre-training for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5754–5764.
[71] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, "Unified language model pre-training for natural language understanding and generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13042–13054.
[72] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," 2018, *arXiv:1806.08730*.
[73] S. Son, S. Hwang, S. Bae, S. J. Park, and J.-H. Choi, "A sequential and intensive weighted language modeling scheme for multi-task learning-based natural language understanding," *Appl. Sci.*, vol. 11, no. 7, p. 3095, Mar. 2021.
[74] M. Mcshane, "Natural language understanding (NLU, not NLP) in cognitive systems," *AI Mag.*, vol. 38, no. 4, pp. 43–56, 2017.
[75] S. Jung, "Semantic vector learning for natural language understanding," *Comput. Speech Lang.*, vol. 56, pp. 130–145, Jul. 2019.
[76] T. Yu, "Gradient surgery for multi-task learning," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2020, pp. 5824–5836.
[77] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," 2018, *arXiv:1804.07461*.
[78] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," 2018, *arXiv:1804.07461*.
[79] A. Wang, "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," in *Proc. 33rd Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 1–15.
[80] K. Narasimhan and M. Hausknecht, "Reading and acting while blindfolded: The need for semantics in text game agents," 2021, *arXiv:2103.13552*.
[81] A. Zweig and D. Weinshall, "Hierarchical regularization cascade for joint learning," in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 2, 2013, pp. 1074–1082.
[82] G. Chen and M. Huang, "Out-of-domain detection for natural language understanding in dialog systems," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1198–1209, 2020.
[83] B. Dhingra, "Gated-attention readers for text comprehension," in *Proc. 55th Annu. Meeting ACL*, Vancouver, BC, Canada, 2017, pp. 1832–1846.
[84] J. Howard, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting ACL*, Melbourne, VIC, Australia, 2018, pp. 328–339.
[85] D. Chen, "A through examination of the CNN/daily mail reading comprehension task," in *Proc. 54th Annu. Meeting ACL*, Berlin, Germany, 2018, pp. 2358–2367.
[86] W. Che, Y. Liu, Y. Wang, B. Zheng, and T. Liu, "Towards better UD parsing," in *Proc. CoNLL*, 2018, pp. 55–64.
[87] D. Kiselev, "An AI using construction grammar to understand text: Parsing improvements," *Int. J. Cogn. Inform. Natural Intell.*, vol. 15, no. 2, pp. 47–61, Apr. 2021.
[88] S. Singh and A. Mahmood, "The NLP cookbook: Modern recipes for transformer based deep learning architectures," *IEEE Access*, vol. 9, pp. 68675–68702, 2021.
[89] C. Montenegro, A. L. Zorrilla, J. M. Olaso, R. Santana, R. Justo, J. A. Lozano, and M. I. Torres, "A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly," *Multimodal Technol. Interact.*, vol. 3, no. 3, p. 52, Jul. 2019.
[90] T. Mairittha, N. Mairittha, and S. Inoue, "Automatic labelled dialogue generation for nursing record systems," *J. Pers. Med.*, vol. 10, no. 3, pp. 1–24, Sep. 2020.
[91] D. Braun, "Evaluating natural language understanding services for conversational question answering systems," in *Proc. SIGDIAL Conf.*, Saarbrucken, Germany, 2017, pp. 174–185.
[92] M. Zubani, L. Sigalini, I. Serina, and A. E. Gerevini, "Evaluating different natural language understanding services in a real business case for the Italian language," *Proc. Comput. Sci.*, vol. 176, pp. 995–1004, Oct. 2020.
[93] A. Jiao, "An intelligent chatbot system based on entity extraction using RASA NLU and neural network," *J. Phys. Conf. Ser.*, vol. 1487, Mar. 2020, Art. no. 012014, doi: 10.1088/1742-6596/1487/1/0120142020.
[94] T. Mayer, "Enriching language models with semantics," in *Proc. 24th Eur. Conf. Artif. Intell.*, Santiago, Spain, Aug. 2020, pp. 1–3.
[95] G. B. Jenset and B. McGillivray, "Enhancing domain-specific supervised natural language intent classification with a top-down selective ensemble model," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 2, pp. 630–640, Apr. 2019.
[96] P. Lauren, "A conversational user interface for stock analysis," in *Proc. Int. Conf. Big Data*, 2019, pp. 5298–5305.
[97] V. Shymko, "Natural language understanding: Methodological conceptualization," *Psycholinguistics*, vol. 25, no. 1, pp. 431–443, Apr. 2019.
[98] E. Tomal and K. D. Forbus, "EA NLU: Practical language understanding for cognitive modelling," in *Proc. 22nd Int. Florida Artif. Intell. Res. Soc. Conf.*, vol. 22, 2009, pp. 117–122.
[99] Z. Li and D. Hoiem, "Learning without forgetting," 2016, *arXiv:1606.09282*.
[100] J.-L. Du, "Predicting garden path sentences based on natural language understanding system," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 11, 2012, doi: 10.14569/IJACSA.2012.031101.
[101] D. Weissenborn, T. Koàiský, and C. Dyer, "Dynamic integration of background knowledge in neural NLU systems," 2017, *arXiv:1706.02596*.
[102] S. Peng and A. Zeldes, "All roads lead to UD: Converting stanford and penn parses to english universal dependencies with multi-layer annotations," in *Proc. Joint Workshop Linguistic Annotation, Multiword Expressions, Construct. Workshop*, 2018, pp. 167–177.
[103] G. Senay, B. Y. Idrissi, and M. Haziza, "VirAAL: Virtual adversarial active learning for NLU," Tech. Rep., May 2020. [Online]. Available: https://arxiv.org/abs/2005.07287V2
[104] B. Wu, B. Wei, J. Liu, K. Wu, and M. Wang, "Faceted text segmentation via multitask learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3846–3857, Sep. 2021, doi: 10.1109/TNNLS.2020.3015996.
[105] M. Dimovski, C. Musat, V. Ilievski, A. Hossmann, and M. Baeriswyl, "Submodularity-inspired data selection for goal-oriented chatbot training based on sentence embeddings," 2018, *arXiv:1802.00757*.
[106] A. Tiwari, T. Saha, S. Saha, S. Sengupta, A. Maitra, R. Ramnani, and P. Bhattacharyya, "A dynamic goal adapted task-oriented dialogue agent," *PLoS ONE*, vol. 16, no. 4, Apr. 2021, Art. no. e0249030.
[107] E. Merdivan, D. Singh, S. Hanke, J. Kropf, A. Holzinger, and M. Geist, "Human annotated dialogues dataset for natural conversational agents," *Appl. Sci.*, vol. 10, no. 3, p. 762, Jan. 2020.
[108] S. Zhu, R. Cao, and K. Yu, "Dual learning for semi-supervised natural language understanding," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1936–1947, 2020.
[109] B. Shi, "Auxiliary task reweighting for minimum-data learning," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2020, pp. 7148–7160.
[110] Y. Ji, C. Tan, S. Martschat, Y. Choi, and N. A. Smith, "Dynamic entity representations in neural language models," 2017, *arXiv:1708.00781*.
[111] L. Liu, X. Ren, J. Shang, X. Gu, J. Peng, and J. Han, "Efficient contextualized representation: Language model pruning for sequence labelling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1215–1225.
[112] R. Caruana, "Multitask learning," *Mach. Learn. J.*, vol. 28, no. 1, pp. 41–75, 1997.

RAHUL MANOHAR SAMANT received the master's degree in information technology from Mumbai University. He is currently pursuing the Ph.D. degree with the Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune. His research interests include conversational AI, deep learning, and language modeling.



MRINAL R. BACHUTE received the M.E. degree in digital electronics and the Ph.D. degree in electronics. She is currently an Associate Professor and an Industry Liaison Officer at the Department of Electronics and Telecommunication Engineering, Symbiosis Institute of Technology, Pune Symbiosis International (Deemed University), Pune, Maharashtra, India. She has teaching experience of 20 years. She has received research funding from the University of Pune and AICTE QIP Grants. Her research interests include digital image processing, machine learning, artificial intelligence, and adaptive signal processing. She has delivered invited talks and expert sessions at the various national and international levels, including at Langara University, Vancouver, Canada, organized by IET Canada at ZE Power Engineering, Vancouver, and IET Trinidad and Tobago. She has worked as a reviewer for conferences and reputed journals, like Springer, Nature, and Elsevier.

SHILPA GITE received the Ph.D. degree in using deep learning for assistive driving in semi-autonomous vehicles from Symbiosis International (Deemed University), Pune, India, in 2019. She is currently working as an Associate Professor with the Computer Science Department, Symbiosis Institute of Technology, Pune. She is also working as an Associate Faculty at the Symbiosis Centre of Applied AI (SCAAI). She has around 13 years of teaching experience. She has published more than 60 research papers in international journals and 25 Scopus indexed international conferences. Her research interests include deep learning, machine learning, medical imaging, and computer vision. She was a recipient of the Best Paper Award at the 11th IEMERA Conference held virtually at Imperial College London, London, in October 2020.

KETAN KOTECHA received the M.Tech. and Ph.D. degrees from IIT Bombay.

He is currently the Head of the Symbiosis Centre for Applied AI (SCAAI), the Director of the Symbiosis Institute of Technology, a CEO of the Symbiosis Centre for Entrepreneurship and Innovation (SCEI), and the Dean of the Faculty of Engineering, Symbiosis International (Deemed University). He has expertise and experience in cutting-edge research and AI and deep learning projects for the last (more than) 25 years. He has published more than 100 articles widely in several excellent peer-reviewed journals on topics ranging from cutting-edge AI, education policies, teaching-learning practices, and AI for all. He was a recipient of the two SPARC projects worth INR 166 lakhs from MHRD Government of India in AI in collaboration with Arizona State University, USA, and The University of Queensland, Australia.

Dr. Kotecha was a recipient of numerous prestigious awards, like the Erasmus+ Faculty Mobility Grant to Poland, the DUO-India Professors Fellowship for research in responsible AI in collaboration with Brunel University, U.K., the LEAP Grant at Cambridge University, U.K., the UKIERI Grant with Aston University, U.K., and a Grant from the Royal Academy of Engineering, U.K., under the Newton Bhabha Fund. He has published three patents and delivered keynote speeches at various national and international forums, including at the Machine Intelligence Laboratory, USA, IIT Bombay under the World Bank Project, the International Indian Science Festival organized by the Department of Science Technology, Government of India, and many more. He is an Academic Editor of the *Peerj Computer Science* journal and an Associate Editor of IEEE Access journal.