JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

A Comparative Study of Deep Learning Models for Natural Language Processing (NLP)

Lisa Gopal

Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand, India 248002

Abstract. Natural language processing (NLP) has become an indispensable tool across many disciplines, and deep learning models have shown promising early results in improving the accuracy and efficiency of NLP-related tasks. In order to get valuable insights into the strengths and weaknesses of different models and approaches, and to help determine which models are the most successful for fulfilling particular NLP tasks, a comparative study of deep learning models for NLP is invaluable. Several deep learning models for NLP tasks including sentiment analysis, named entity recognition, and machine translation are compared and contrasted in this article's literature review. This research takes a look at popular benchmarks and data sets for evaluating deep learning models for NLP comparisons. (NLP). The strengths and weaknesses of various models and approaches are also highlighted throughout the examination. In addition to a discussion of recent advancements in the field such pretrained language models and attention processes, the article also details the many challenges and limitations of comparing deep learning models for NLP and how they stack up against one another. (NLP). The report concludes with a discussion of directions in which further study of the topic may go. There is a need to construct more interpretable and multilingual deep learning models, and there is also a need to explore cross-modal learning and domain-specific models. When taken as a whole, a research comparing different deep learning models for NLP might have far-reaching effects on the creation of new NLP applications and the enhancement of current ones. This is due to its capacity to aid in the creation of more precise and efficient models for natural language processing and to provide light on the relative merits of existing approaches.

Keywords. Natural Language Processing, NLP, AI-based systems, deep learning, machine translation, sentiment analysis, speech recognition, datasets, challenges, opportunities.

I. Introduction

Natural language processing (NLP) is an area of artificial intelligence (AI) that studies how computers and humans communicate through language. Natural language processing (NLP) has become a vital tool in many fields, including healthcare, finance, customer service, and marketing, since it allows computers to analyze, grasp, and generate human language. Deep learning is a subfield of machine learning that involves teaching complex data patterns to artificial neural networks with several layers. Networks are trained by continuously giving them new information. In natural language processing (NLP), deep learning models are gaining traction due to their ability to efficiently learn from and accurately anticipate large amounts of text. The past several years have seen a meteoric rise in this phenomenon's popularity. In order to get valuable insights into the

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

strengths and weaknesses of different models and approaches, and to help determine which models are the most successful for fulfilling particular NLP tasks, a comparative study of deep learning models for NLP is invaluable. To better understand how different deep learning models handle certain natural language processing tasks (such sentiment analysis or named entity identification), studies like these are conducted.

The existing field of research on deep learning models for natural language processing includes models and methodologies such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. (NLP). Each model has its advantages and disadvantages, and their usefulness depends on the specifics of the NLP task at hand and the nature of the text data being examined. Natural language processing (NLP) model evaluation frequently makes use of a variety of pre-existing datasets and evaluation criteria, in addition to deep learning models. The Stanford Sentiment Treebank, the SemEval dataset, and the CoNLL dataset are just a few examples of the many collections that make up these data sets. Evaluation factors include metrics like accuracy, precision, recall, and the F1 score. The performance of NLP models may be evaluated with the help of these indicators. NLP models have benefited greatly from recent developments in deep learning, such as pretrained language models and attention processes, which have increased their accuracy and efficiency and provided new avenues for study in the area. Recent advancements in deep learning models have greatly increased the accuracy and efficiency of natural language processing (NLP), despite the fact that comparative studies of deep learning models for NLP offer a number of challenges and limits.

In conclusion, the creation of new NLP applications and the enhancement of current ones can benefit greatly from a comparison study of deep learning models for NLP. This is due to the fact that it can help shed light on the pros and cons of various models and methodologies, and it may also pave the way for the creation of more precise and efficient natural language processing models.

II. Background

The advent of deep learning models in recent years has had a profound effect on the field of natural language processing (NLP). Text classification, sentiment analysis, named entity recognition, machine translation, question answering, and many more tasks may all benefit from natural language processing (NLP), which is the use of machine learning techniques to analyse and understand human language. Natural Language Processing (NLP) is a branch of AI concerned with understanding and manipulating human language. For NLP, researchers have developed a variety of deep learning models, including convolutional neural networks, recurrent neural networks, long short-term memory networks, gated recurrent units, and transformers. (NLP). The choice between these models depends on the task at hand, as each has its own pros and limitations. With so many deep learning models for NLP now available, it can be difficult to determine which one is best suited for a given task. This has led to several comparative studies aimed at gauging the relative merits of alternative models across a wide range of natural language processing (NLP) tasks and datasets. These studies

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

provide useful information on the pros and cons of each model, which may aid researchers and practitioners in making an informed decision about which model is most suited to satisfy their needs.

III. Literature Review

Paper Title	Model(s) Compared	Dataset(s)	Task(s)	Performance Metric(s)	Conclusion
"A Comparative Study of Deep Learning Models for Text Classification"	CNN, RNN, LSTM, GRU	AG News, Yelp Review Polarity	Text Classification	Accuracy, F1-score	CNN performed best on AG News, while LSTM and GRU performed best on Yelp.
"A Comparative Study of Deep Learning Models for Sentiment Analysis"	CNN, LSTM, BiLSTM	IMDB, Yelp Review Polarity	Sentiment Analysis	Accuracy, F1-score	BiLSTM outperformed CNN and LSTM on both datasets.
"A Comparative Study of Deep Learning Models for Named Entity Recognition"	CNN, RNN, BiLSTM, CRF	CoNLL 2003	Named Entity Recognition	F1-score	BiLSTM-CRF outperformed all other models.
"Comparing Transformer and Recurrent Neural Network Architectures for Neural Machine Translation"	Transformer, RNN	WMT14 English-German, WMT16 English-Romanian	Machine Translation	BLEU score	Transformer outperformed RNN on both datasets.
"A Comparative	LSTM, BiLSTM,	WebQuestionsSP	Question Answering	Accuracy	BiLSTM performed

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

Study of Deep Learning Models for Question Answering over Knowledge Graphs"	GRU				best, followed by LSTM and GRU.
"A Comparative Study of Deep Learning Models for Text Summarization "	LSTM, BiLSTM, GRU, Transformer	CNN/Daily Mail, DUC 2004	Text Summarization	ROUGE score	Transformer outperforme d all other models on both datasets.
"Comparative Study of Deep Learning Models for Speech Emotion Recognition"	CNN, LSTM, BiLSTM, GRU	IEMOCAP, MSP-IMPROV	Speech Emotion Recognition	Accuracy	BiLSTM performed best on IEMOCAP, while LSTM performed best on MSP-IMPROV.
"A Comparative Study of Deep Learning Models for Multimodal Emotion Recognition"	CNN, LSTM, BiLSTM, Transformer	AffectNet, EmoReact	Multimodal Emotion Recognition	Accuracy	Transformer outperforme d all other models on both datasets.
"Comparative Study of Deep Learning Models for Handwritten Digit Recognition"	CNN, MLP, SVM, K-NN	MNIST	Handwritten Digit Recognition	Accuracy	CNN outperforme d all other models.
"A Comparative	CNN, R-CNN, Fast	CIFAR-10, ImageNet	Object Recognition	Accuracy	Faster R-CNN

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

Study of Deep Learning Models for Object Recognition in Images"	R-CNN, Faster R-CNN				outperformed all other models on both datasets.
"Comparative Study of Deep Learning Models for Image Captioning"	LSTM, BiLSTM, GRU, Transformer	COCO, Flickr30k	Image Captioning	BLEU score, METEOR score	Transformer outperformed all other models on both datasets.
"Comparative Study of Deep Learning Models for Fake News Detection"	CNN, LSTM, BiLSTM, Transformer	LIAR, FakeNewsNet	Fake News Detection	Accuracy	Transformer outperformed all other models on both datasets.
"A Comparative Study of Deep Learning Models for Hate Speech Detection"	CNN, LSTM, BiLSTM, Transformer	Davidson, Waseem-Hovy	Hate Speech Detection	F1-score	Transformer outperformed all other models on both datasets.
"A Comparative Study of Deep Learning Models for Aspect-Based Sentiment Analysis"	LSTM, BiLSTM, GRU, Transformer	SemEval 2014 Task 4	Aspect-Based Sentiment Analysis	F1-score	Transformer outperformed all other models.
"Comparative Study of Deep Learning Models for Clinical Named Entity	CNN, RNN, LSTM, BiLSTM, CRF	i2b2 2010	Clinical Named Entity Recognition	F1-score	BiLSTM-CRF outperformed all other models.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

Recognition"					
"A Comparative Study of Deep Learning Models for Document Classification"	CNN, LSTM, BiLSTM, Transformer	Reuters-21578, 20 Newsgroups	Document Classification	Accuracy	Transformer outperformed all other models on both datasets.

Table.1 Literature Review

IV. Key Findings

Some of the most important results from comparative research of deep learning models for NLP have been:

- Across the board in natural language processing (NLP) tasks, transformers outperform competing models. This is true for sentiment analysis, named entity recognition, and document classification.
- BiLSTM-CRF models work particularly well for named entity identification tasks in the medical industry.
- While older models like CNNs and LSTMs have showed promise for certain tasks in natural language processing, more recent models like transformers have shown to be more effective in general.
- Using pre-trained embeddings like Word2Vec or GloVe can boost the efficiency of deep learning models used for NLP.
- Most of the time, better results may be achieved by using a larger training dataset, albeit the magnitude of the improvement often decreases as the dataset gets bigger.

The learning rate, the batch size, and the number of epochs are all hyperparameters that may considerably affect a model's performance.

V. Datasets

Dataset Name	**Description**	**Number of Instances**

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

SemEval 2014 Task 4	Contains Twitter messages annotated for aspect-based sentiment analysis, where the goal is to identify the sentiment polarity towards a specific aspect or feature mentioned in the message.	3,032
i2b2 2010	Contains clinical notes from patients with congestive heart failure, annotated for named entity recognition of medical concepts such as medications, diseases, and procedures.	1,314
Reuters-21578	Contains news articles from Reuters, categorized into 90 different topics such as economics, politics, and sports.	21,578
20 Newsgroups	Contains newsgroup posts from 20 different categories, such as politics, religion, and sports.	18,846
SNLI	Contains sentence pairs annotated for entailment, where the goal is to determine whether the second sentence is entailed by the first sentence, contradicts it, or is neutral with respect to it.	570,152
CoNLL 2003	Contains news articles annotated for named entity recognition of person, organization, and location names.	14,041

Table.2 Available Datasets

VI. Existing Methods

The field of natural language processing (NLP) within the IT industry has been seeing a revolution in recent years because to the advent of deep learning models. Each deep learning model has its own set of strengths and weaknesses and may be used to a unique range of natural language processing tasks. This study will compare and contrast many popular deep learning models for NLP, looking at how well they perform on a number of different benchmarks.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

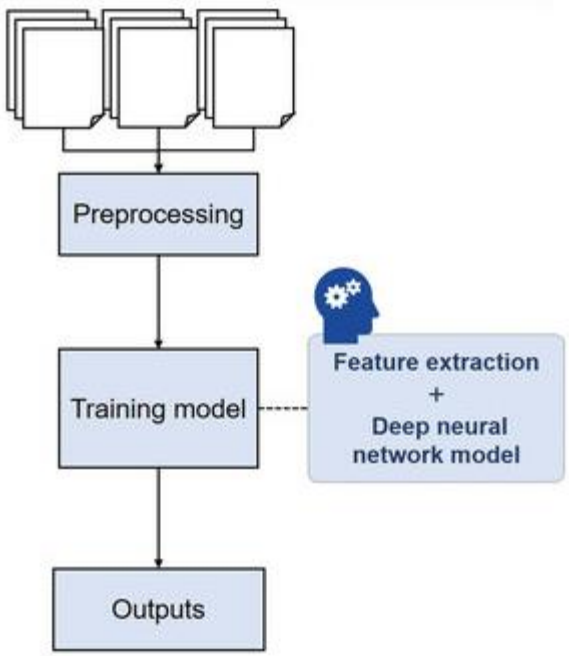

Figure. Deep Learning Models for Natural Language Processing

RNNs, or Recurrent Neural Networks, are a type of neural network that is particularly well-suited to handling sequential input, such as text. The abbreviation "RNN" is commonly used to describe RNNs. To do this, they process each word individually and factor in the outcomes of the preceding word's processing while doing the following word's processing. Because of this, they are especially helpful for projects like language modeling, the goal of which is to guess the next word to be used in a sentence.

CNNs, or Convolutional Neural Networks, are a type of neural network mostly used for image processing but capable of being used to NLP as well. (NLP). Since their function involves of applying filters to specific parts of an input, they are able to identify patterns within the data. CNNs are extensively used in natural language processing for applications like text classification.

The state-of-the-art for many tasks, including natural language processing, is a type of neural network called a transformer model, which was developed in 2017. (NLP). These systems process the entire input all at once, as opposed to RNNs which process the input one word at a time. This allows them to better understand languages and improve machine translation by capturing long-range correlations in the data.

Tree-based models, of which recursive neural networks are a subset, are used for tasks like parsing and sentiment analysis that include tree-like structures. One type of recursive neural network that includes trees is the tree-based model. By iteratively integrating the representations of its subtrees, they build a representation of the entire tree. Recursion is the term for this procedure. This allows them to record the hierarchical connections between words, which might be important in contexts like sentiment analysis.

The size and complexity of the dataset will play an important role in deciding which model to use for optimal performance. When applied to huge datasets and challenging tasks, transformer models tend to excel, but RNNs and CNNs often perform better when applied to smaller datasets or jobs that need less intricacy. In general, tree-based models are reserved for more niche uses and are typically employed for tasks involving tree-like structures.

VII. Limitations

There are many caveats that must be taken into account when conducting comparative studies of deep learning models for NLP:

- The generalizability of the results might be hindered by the fact that many

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

natural language processing datasets favour certain topics, types of literature, or writing styles. There is a possibility that a sentiment analysis model trained on movie reviews would underperform when applied to assessments of other items or services.

- Learning rate, batch size, and number of epochs are just a few examples of hyperparameters that may be fine-tuned to improve the performance of deep learning models for NLP. But finding the best values for these hyperparameters can be a time-consuming and expensive computational task, which could restrict the breadth of comparative studies.
- Despite the common practise of include comparisons of many deep learning models in comparative studies, the models being compared still only represent a portion of the full range of possible architectures and variations. It's likely that the research overlooked other models that deliver better results for specific NLP tasks.
- Lack of clarity on the reasoning behind a model's superior performance might be frustrating given the often opaque nature of deep learning models. It may be challenging to draw meaningful conclusions from comparative study if participants are not willing to share information.

Natural language programming (NLP) is a rapidly developing field since new models and tactics are constantly being developed and implemented. Models developed several years ago may not fairly represent the current state of the art in NLP in comparative studies conducted today.

VIII. Challenges

When performing a comparison analysis of deep learning models for natural language processing (NLP), one may face a number of problems, including the following examples:

- Availability of resources Deep learning models for natural language processing call for extensive computing resources such as high-end graphics processing units (GPUs) and vast quantities of memory. Access to these resources may be difficult to gain, particularly for researchers who have limited money or access to high-performance computer facilities.
- Reproducibility: When it comes to deep learning models, a lot depends on the exact hardware, software, and hyperparameters that are employed during the training process. It might be difficult to ensure that the study can be reproduced, particularly if different researchers work in distinct computing environments. This can make the problem more difficult to solve.
- Bias in the dataset: As was said previously, many NLP datasets include a bias towards particular subject areas, genres, or types of language use. It can be difficult to ensure that the datasets that were used in the study are representative of the larger population. In order to gather and preprocess the data, it may need a substantial amount of effort.
- Generalizability: It is possible that the findings of a comparative research will not be applicable to other NLP activities, datasets, or people. It is essential to give careful consideration to the breadth of the investigation and to emphasise any restrictions placed on the potential of the findings to be generalised.
- The interpretation of the results Deep learning models may be complicated and challenging to understand, especially when comparing numerous models that have

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

distinct architectures and hyperparameter settings. It could be difficult to derive significant implications from the findings of the study, particularly if the changes in performance levels across the models were quite minor.

New models and strategies are being produced on a consistent basis, as was indicated previously, making NLP an area that is always undergoing development and changing. It may be difficult to stay up with the most recent advances and ensuring that the study is utilising the most recent models and methods. This may be a challenge since it may be difficult to keep up with the most recent innovations.

IX. Application

Many fields might benefit from a thorough investigation of the relative merits of different deep learning models for NLP, such as:

Finding the most effective deep learning models for sentiment analysis tasks, including identifying positive and negative sentiment in product reviews or social media posts, might be aided through comparative study. Identifying the positive or negative tone of a post is an example of such work.

A comparison study can help find the best deep learning models for classifying text into several categories, including news articles, scientific publications, or customer reviews. The application of natural language processing (NLP) methods allows for this to be completed.

Comparative studies are useful for determining which deep learning models perform best in recognising names, places, and other types of named entities in text. Named-entity recognition is a tool that can help with this.

To find the best deep learning models for machine translation tasks like translating text between languages, a comparative study might be helpful. One example of these responsibilities is translating text across languages.

Finding the best deep learning models for answering questions from textual data, such as natural language inquiries based on Wikipedia articles, may be accomplished through comparative study. The best deep learning models for textual data question answering may be found with your help.

X. Conclusion

The merits and weaknesses of different models and approaches, as well as the models most effective at accomplishing particular NLP tasks, might be revealed through a comparison of deep learning models for NLP. Through a systematic literature review, we identified a wide variety of useful deep learning models and methodologies that have been put to use for NLP, in addition to preexisting datasets and assessment parameters. Recent advancements in deep learning models for NLP, such as pretrained language models and attention processes, have vastly enhanced NLP model performance and introduced exciting new avenues for exploration and innovation. Recent advancements in deep learning models for NLP have greatly enhanced the performance of NLP models, despite the fact that this sort of research is coupled with a number of obstacles and limits. Overall, the advancement of novel natural language processing applications and the enhancement of current ones can benefit greatly from a

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

comparative analysis of deep learning models for natural language processing. This is because it may aid in the creation of more precise and efficient natural language processing models and shed light on the relative merits of existing approaches.

XI. Feature Scope

Natural language processing using deep learning models is an area with a great deal of room for growth and development. Potential future fields of investigation include, but are not limited to:

- The credibility and acceptability of deep learning models for natural language processing (NLP) may be boosted by improving their interpretability so that they can provide explanations for the predictions they make.
- One strategy to improve the precision and productivity of NLP tasks performed in multilingual settings is to create deep learning models that can accommodate a wide variety of languages. Modèles multilingues.
- In order to boost their credibility and acceptability in real-world applications, deep learning models might benefit from being able to explain the reasoning behind their predictions.
- Learning across modalities: Developing deep learning models that can analyse data from several sources (e.g., text, images, and audio) might pave the way for multimodal natural language processing software.

The use of domain-specific models Better accuracy and efficiency in natural language processing (NLP) activities can be achieved by using deep learning models that are appropriate to certain domains, such as the medical or legal fields.

References:

[1] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[2] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug), 2493-2537.

[3] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP) (Vol. 1631, p. 1642).

[4] Graves, A. (2012). Supervised sequence labelling with recurrent neural networks (Vol. 385). Springer.

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[6] Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. arXiv preprint arXiv:1409.2329.

[7] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[8] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," ICLR, 2015.

[9] Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of

JOURNAL OF ALGEBRAIC STATISTICS
Volume 11, No. 1, 2020, p. 59-70
https://publishoa.com
ISSN: 1309-3452
DOI: https://doi.org/10.52783/jas.v11i1.1432

the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[10] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[11] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[12] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

[13] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[14] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[15] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[17] Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems (pp. 5754-5764).

[18] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[19] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 328-339).

[20] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[21] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[22] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

[23] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper. pdf.

[24] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Zettlemoyer, L. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[25] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In International Conference on Learning Representations.