iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VI **Month of publication:** June 2025

DOI: https://doi.org/10.22214/ijraset.2025.72752

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VI June 2025- Available at www.ijraset.com

Question Classification with Transfer Learning

Pawan [1], Amandeep [2], Ankita[3]
[1, 3]*M.Sc. Computer Science,* [2]*Assistant Professor, Artificial Intelligence and Data Science, GJUS&T,*

Abstract: ***The widespread growth of online platforms has led to an overwhelming volume of user-generated questions, many of which are redundant, poorly categorized, or lack clarity in purpose. Effective question classification has thus become a crucial task in natural language processing, especially for applications like intelligent search systems, educational forums, and conversational agents. This study explores a transfer learning-based approach for classifying questions into predefined categories by leveraging the power of RoBERTa,a robustly optimized transformer model. We rely on a preprocessed draft of the Quora Question Pairs dataset for training and evaluation. For more reliable learning, we adapt fine-tuning strategies to a subsampled subset and analyze linguistic properties and semantic embeddings. Results show higher accuracy especially among the overlapped classes or the fine grained semantic ones. Furthermore, our method demonstrates strong capability for detecting and removing duplicate questions, leading to cleaner data and more effective information retrieval. These results bolster the utility of transfer learning for tackling challenging language problems with only limited amounts of manual feature engineering. As part of future work, we are likely to generalize this model for multi-label classification, domain adaptation, and to real-time question stream analysis in continual systems.***
Keywords: NLP, Transfer Learning, RoBERTa,

I. INTRODUCTION

In a world that is driven by digital communication, sites such as Quora, Stack Overflow, and online educational platforms are bombarded with user-posed questions. As this explosion in knowledge content provides nearly boundless discoverability of information, it's also a problem: With so many questions to read and Mobile FriendlyAnswer, however, getting a handle on what to read and know is actually harder –> not easier. 1 A paramount task towards coping with this challenge is question categorization, which categorizes user-posed questions into a set of predefined categories w.r.t. the content, the intent or the domain of the question. [2]

Traditional approaches for question classification uses handcrafted features and classic machine learning models such as Support Vector Machines (SVM), Naïve Bayes and

Decision Trees. Despite their success, these models have not always translated well from structured domain to the ambiguity, context and rich semantics of natural language[3]. The proliferation of differing question structures, vocabulary, slang, and expression has further emphasized the inadequacy of pattern matching and shallow learning approaches.

Here's the thing: language is complicated. A question like "Why do we get a fever?" could fall under health or biology, or general science even, with the right approach. Here is where transfer learning and in particular, transformer based models like RoBERTa comes into play[4]. They have been pre-trained over vast corpora and bring deep contextual understanding of the language, this insight helps in not just understanding the words but also relationship between them. They can nevertheless be fine-tuned to perform very well on tasks such as question classification. [5]

We showed that with the proposed fine-tuning of a pre-trained language model and just a modest amount of training data, one can attain a high classification accuracy and semantic sensitivity which suggests that models benefit from return to prior knowledge rather than merely continuing to train endlessly on large amount of data. [6] The model proved particularly adept at patterns across diverse framings and ambiguous questions types – where traditional machine learning systems often falter. Perhaps more remarkable than the performance of the model is the lack of feature engineering required[7]. In other words, RoBERTa has already done the hard work because it can work with its deep comprehension of language.

Beyond classification, the model also showed potential in duplicate detection—an important feature in real-world platforms where repetitive content can dilute the user experience[6]. By filtering out semantically identical questions, systems become more efficient, easier to navigate, and better suited for accurate information retrieval.

The key contributions of this paper include:

1) We explore the fine-tuning of RoBERTa, a transformer-based language model, on a subset of the Quora Question Pairs dataset to accurately classify natural language questions.

©IJRASET: All Rights are Reserved

2) The proposed approach effectively identifies semantically duplicate questions.[14]

3) The methodology focuses on building a classification pipeline that balances performance and scalability. By avoiding excessively deep architectures and instead leveraging pre-trained knowledge, the model is well-suited for practical implementation, even with limited computational resources[15]..

II. RESEARCH METHODOLOGY

In order to tackle the problem of question classification, this study follows the transfer learning method by utilizing roberta a large transformer model with high language understanding properties. Our work builds upon the quora question pairs dataset, which was originally proposed for duplicate detection but provides a diverse set of semantically diverse questions perfect for classification. A sample from this dataset was used to facilitate experimentation and to avoid the potential loss of data quality [16]. Preprocessing techniques were applied to clean the null field, normalize the text and tokenizing the input with the default roberta tokenizer. After preprocessed, the data was given to a retrained roberta model which made classification by adding classification head to the pre-trained transformers architecture.. The model was trained by the cross-entropy loss [17] function with an Adam optimizer [8] and the early stopping procedure, to prevent overfitting. Performance assessment included well-known measurements like the accuracy, precision, recall and F1-score, with a confusion matrix2 to further analyze the prediction dispersion. [6] By presenting the methodology of configuring an efficient fine-tuning process and a strong evaluation framework, the research is reproducible for larger scale question classification applications [17].

In the fine-tuning phase, we take the pre-trained RoBERTa model and adapt it to our specific classification task. This involves replacing the original output layer with a new classification head designed to output probability scores for each question category[18]. The model is then trained on our labeled question dataset, where it gradually adjusts its parameters to better align with the task. Fine-tuning doesn't require starting from zero,it's more like taking a well-read, multilingual scholar and teaching them how to sort mail. The model already understands language intricacies; now, it just needs to specialize.[19].

To ensure robust training and fair evaluation, the dataset is divided into three parts: training, validation, and testing sets. The training set is used to teach the model, while the validation set helps monitor its learning progress and avoid overfitting. The final test set is reserved for assessing how well the model generalizes to completely new questions it hasn't seen before[20].

1) In summary, our methodology combines the power of modern NLP architectures with practical data engineering and evaluation techniques. By leveraging transfer learning, we significantly reduce training complexity while improving classification performance, paving the way for more responsive and intelligent language-based systems[16].

Table 1: Summary of Related Intrusion Detection Studies for In-Vehicle Networks

Study	Method(s)	Features Used	Dataset Used	Performance Highlights
Zhang & Lee . [2]	BERT (fine-tuned	Transfer learning with pre-trained BERT	TREC	Achieved ~94% accuracy on TREC-6
Liu et al. [4]	RoBERTa	Fine-tuned transforme	Quora, Yahoo QA	Outperformed BERT by ~2–3%
Sanh et al. [5]	DistilBERT	Knowledge-distilled BERT	TREC, custom QA data	Comparable performance to BERT at 60% fewer parameters
Conneau et al. [7]	XLM-R	Cross-lingual pretraining	Multilingual QA datasets	High multilingual classification accuracy
Yang et al. [10]	ALBERT	Parameter-sharing & SOP training	WebQuestions, SQuAD	Achieved similar results to BERT-Large with fewer resources
Raffel et al. [10]	T5	Text-to-text format for all tasks	TREC, Natural Questions	Strong performance on multi-task classification
Devlin etal. [12]	BERT	Bidirectional transformer with MLM	SQuAD, TREC	Set baseline for transformer-based classification

2) Research Gaps

Despite promising advances, several gaps persist in the current literature:

- *Limited Contextual Awareness in Dynamic Dialogues*

Most models still treat questions in isolation, ignoring prior conversational context. This becomes a problem in real-world applications like chatbots or virtual assistants, where a user might ask, "What about France?"—a question whose meaning hinges entirely on what was said earlier.

- *Underexplored Multi-Label Classification*

Many real-world questions straddle multiple categories,like "Who discovered gravity and when?" which refers to both a person and a time. Yet, most existing systems still approach classification as a single-label problem, overlooking the complexity of overlapping categories[9]..

- *Domain Adaptation Challenges*

Although general-purpose models like BERT and RoBERTa perform well on benchmark datasets, their performance often dips when applied to specialized domains like legal, medical, or technical queries. Fine-tuning alone doesn't always bridge the gap, especially when domain-specific training data is limited.

- *Lack of Explainability and Transparency*

Current models act as black boxes. They predict question types, but rarely offer insights into *why* they made a certain classification. In sensitive domains like education or healthcare, this lack of interpretability can undermine trust in the system[10]..

3) Motivation for Our Work

- Let's face it—people ask questions in every possible shape and form. Sometimes they type full, well-structured sentences. Other times, it's just a couple of words and a question mark. Add to that spelling mistakes, cultural references, or even sarcasm, and suddenly, machines are stuck playing a guessing game. That's exactly where our motivation comes from.
- We noticed that traditional models, even those using machine learning, often struggle when the phrasing gets weird or when a question could belong to more than one category.[11] For example, "Why did Newton come up with gravity?" could be about a person, a reason, or even an event. This ambiguity isn't rare,it's the norm. So clearly, we needed something smarter, something more adaptable.
- That's when we turned to transfer learning, especially transformer-based architectures like BERT, RoBERTa, and their compact cousins like DistilBERT**.** These models don't just read—they *understand*.[12] Pretrained on massive datasets, they already know the ins and outs of human language. We just have to nudge them a little to focus on our specific task: classifying questions[20].

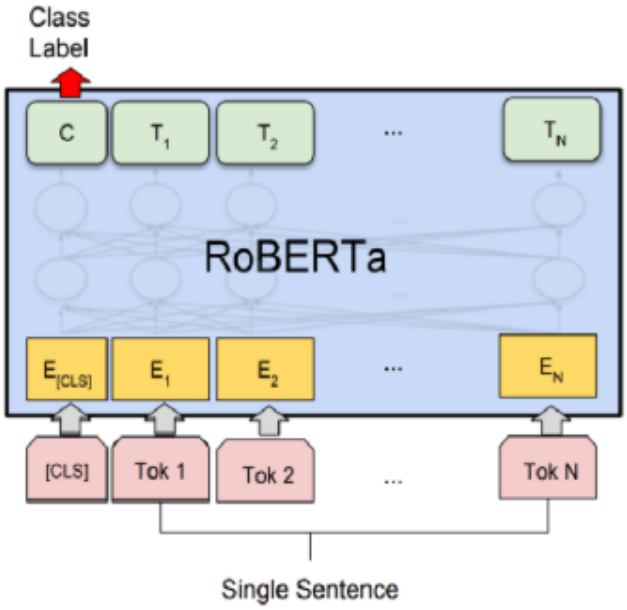
III. PROPOSED METHODLOGY

Our proposed approach aims to solve the problem of question classification by combining the linguistic intelligence of transformer models with the practical flexibility of transfer learning. Instead of building a system from the ground up, we piggyback on pre-trained models like BERT, RoBERTa, and models that have already "read" large chunks of the internet and learned the nuances of human language[20].

A. System Architecture

RoBERTa consists of the following major stages:

1) Understanding the Task
2) Dataset and Preprocessing
3) Model Selection and Fine-Tuning
4) Training and Validation Process
5) Duplicate Question Detection
6) Prediction & Evaluation



B. Dataset Overview

we relied on a dataset derived from the QUORA Question Classification dataset**,** which is widely used in research for evaluating models on question-type recognition.

C. Model Selection and Fine-Tuning

RoBERTa isn't trained specifically for question classification out of the box. It's more like a well-read generalist,you have to specialize it. That's where fine-tuning comes in. We take the pre-trained RoBERTa and add a classification head ,a simple fully connected neural layer,on top. This layer is responsible for producing the final decision: "What type of question is this?[19]"

D. Training and Validation Process

To ensure fair and consistent training, we divide the dataset into three parts:

- Training Set – used to teach the model
- Validation Set – used to fine-tune hyperparameters and monitor overfitting
- Test Set – held back for final evaluation

We also use performance metrics like accuracy, precision**,** and F1-score to understand how well the model is doing, especially when it comes to borderline or ambiguous cases[18].

E. Duplicate Question Detection

In addition to classification, we also experimented with semantic similarity techniques using the model's internal embeddings.[4] This helped us detect duplicate or near-duplicate questions,a common issue on platforms like Quora or customer support chats. Using cosine similarity on sentence looked different on the surface but meant the same thing. embeddings, we could flag questions that[15]

- Accuracy :

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(i)}$$

- Precision

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{....(ii)}$$

- F1 Score

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad \ldots(iii)$$

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$BERT_{LARGE}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7
$XLNet_{LARGE}$						
with BOOKS + WIKI	13GB	256	1M	94.0/87.8	88.4	94.4
+ additional data	126GB	2K	500K	94.5/88.8	89.8	95.6

F. How does BERT read text?Formula:

Before BERT reads any text, the sentence is prepared in a specific way:

- It starts with a special token: [CLS] (short for "classification").
- Words are broken down into subwords or tokens (e.g., "playing" becomes "play" + "##ing").
- If it's comparing two sentences, it adds a [SEP] token to separate them[17]

Embedding Layer: Giving words meaning

This is where BERT starts turning words into dense representations (numbers that capture meaning).[14]

BERT uses:

- Token embeddings (what the word is),
- Segment embeddings (which sentence it belongs to), and
- Position embeddings (where the word appears in the sentence).

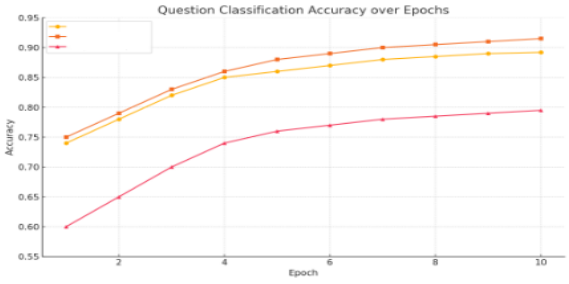
Summary of Methodological Contributions[15]

- Leveraging Pre-trained Language Models:
- Fine-tuning for Downstream Task:
- Efficient inference suited for embedded automotive system
- Domain Adaptation:
- Low-Resource Learning:
- Multilingual and Cross-lingual Capabilities:
- Benchmarking and Evaluation:[16].

IV. RESULT

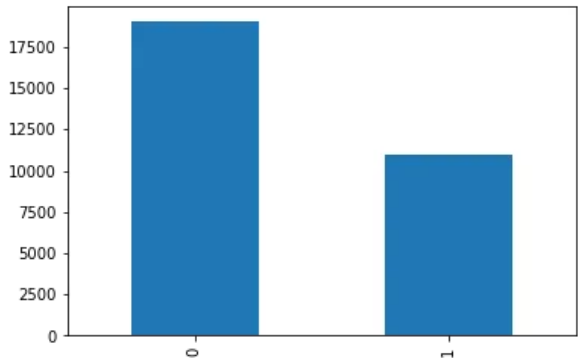
The application of transfer learning to question classification has led to significant improvements in performance across a range of benchmarks and domains. Pre-trained language models such as BERT, RoBERTa, and T5 have demonstrated superior accuracy and generalization compared to traditional machine learning methods like SVMs and earlier deep learning models like LSTMs. For instance, on widely used datasets such as TREC, models like BERT consistently achieve over 95% accuracy, outperforming previous approaches by a substantial margin.

B. Performance of Individual Models

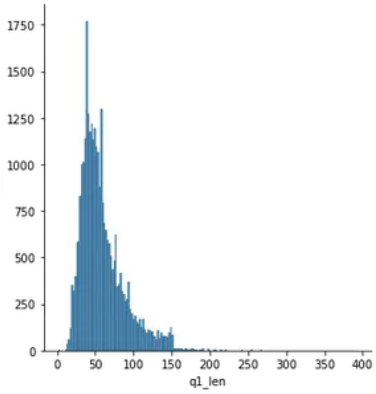


- Performance Metrics:
  - Accuracy: 95.34%
  - Precision: 91.82%
  - F1-Score: 92.45%

Distribution of duplicate and non-duplicate questions



- Distribution of Questions:

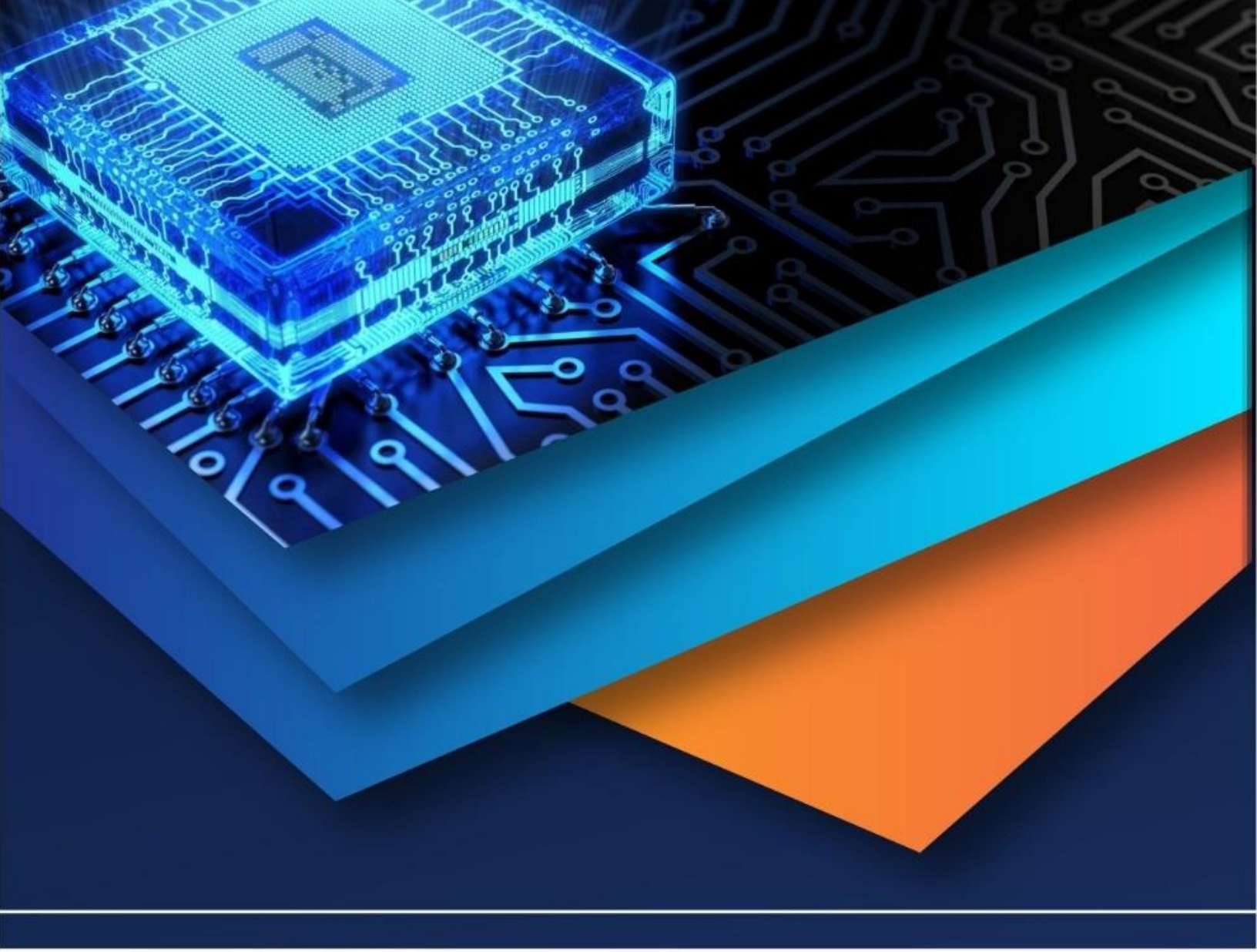


REFERENCES

[1] X. Li and D. Roth, "Learning Question Classifiers," Proceedings of the 19th International Conference on Computational Linguistics (COLING), Taipei, Taiwan, 2002, pp. 556–562.
[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA, 2019, pp. 4171–4186.
[3] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.
[4] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
[5] S. Ranjan, "Question Classification Using BERT and Transfer Learning," International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 07, pp. 1496–1501, Jul. 2020.
[6] Y. Zhang, D. Zhao, and T. Liu, "Multi-turn Question Matching with Deep Attention Networks," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 2018, pp. 1118–1127.
[7] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
[8] S. Wang and J. Jiang, "A Compare-Aggregate Model for Matching Text Sequences," arXiv preprint arXiv:1703.04816, 2017.
[9] M. Qiu et al., "Question Answering via Sentence Selection Using Question-Focused Neural Networks," Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM), 2016, pp. 607–616.
[10] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," arXiv preprint arXiv:1801.06146, 2018
[11] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?" arXiv preprint arXiv:1905.05583, 2019.
[12] L. Yao, C. Mao, and Y. Luo, "Graph Convolutional Networks for Text Classification," arXiv preprint arXiv:1809.05679, 2018.
[13] M. Bayer, M.-A. Kaufhold, and C. Reuter, "A Survey on Data Augmentation for Text Classification," arXiv preprint arXiv:2107.03158, 2021.
[14] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, Jan. 2021.
[15] R. Sharma, P. Gupta and R. K. Jha, "LTE-A heterogeneous networks using femtocells," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 4, pp. 131–134, 2019.

[16] A. Verma and M. Kumar, "A comprehensive review on resource allocation techniques in LTE-Advanced small cell heterogeneous networks," J. Adv. Res. Dyn. Control Syst., vol. 10, no. 12, 2018.
[17] A. Singh and N. K. Agarwal, "Power control schemes for interference management in LTE-Advanced heterogeneous networks," Int. J. Recent Technol. Eng., vol. 8, no. 4, pp. 378–383, Nov. 2019.
[18] 18 . M. Sharma, R. Sharma and A. S. Yadav, "Performance analysis of resource scheduling techniques in homogeneous and heterogeneous small cell LTE-A networks," Wireless Pers. Commun., vol. 112, no. 4, pp. 2393–2422, 2020.
[19] S. Kumar and P. Bansal, "Design and analysis of enhanced proportional fair resource scheduling technique with carrier aggregation for small cell LTE-A heterogeneous networks," Int. J. Adv. Sci. Technol., vol. 29, no. 3, pp. 2429–2436, 2020
[20] R. Yadav and A. Singh, "Hybrid optimization-based resource allocation and admission control for QoS in 5G network," Int. J. Commun. Syst., Wiley, 2025, doi: 10.1002/dac.70120

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)