IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 11, November 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

© 2024 IJIRCCE | Volume 12, Issue 11, November 2024| DOI: 10.15680/IJIRCCE.2024.1211116

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Project Proposal Similarity Detection using NLP and AI

Aditya Shyamanand Mishra, Rohitkumar Brahmdeo Pandey, Shaikh Didar Abbas, Mohammed Ahsan Ansari

Prof. Shiburaj Pappu, Prof. Anupam Choudhary

Dept. of Computer Engineering, Rizvi College of Engineering, Mumbai, India

Dept. of Computer Engineering, Rizvi College of Engineering, Mumbai, India

Dept. of Computer Engineering, Rizvi College of Engineering, Mumbai, India

Dept. of Computer Engineering, Rizvi College of Engineering, Mumbai, India

Dean, Dept. of Computer Engineering, Rizvi College of Engineering, Mumbai, India

HoD, Dept. of Computer Engineering, Rizvi College of Engineering, Mumbai, India

ABSTRACT: In recent years, educational institutions have observed a growing concern over the repetition of student project proposals, which diminishes the originality and innovation of submitted ideas [1]. This project aims to address this issue by implementing an automated system that compares new proposals with previously submitted projects. The system analyzes the title and abstract of each submission using Natural Language Processing (NLP) techniques and Artificial Intelligence (AI) models to compute similarity scores [2]. If the similarity between a new proposal and existing projects exceeds a predefined threshold, the system automatically rejects the submission to ensure the uniqueness of projects. The methodology not only assists in maintaining the integrity of project submissions but also can be adapted for use in the patent application process to enhance efficiency. By leveraging AI and NLP, this system holds significant potential for preventing idea duplication in academia and beyond.

KEYWORDS: Project Similarity Detection, Natural Language Processing, Artificial Intelligence, Academic Projects, Idea Duplication, Patent Application Process

I. INTRODUCTION

This project focuses on identifying similarities between newly submitted student project proposals and previously stored ones by leveraging **Natural Language Processing (NLP)** and **AI techniques** [2]. The system compares the **title** and **abstract** of each proposal to detect duplicates or highly similar ideas, ensuring originality in academicprojects and preventing repetition. Additionally, the system has potential future applications in areas like **patentapplication processes**, where novelty plays a critical role. The primary aim of this project is to automate the manual workload of faculty members. Traditionally, teachers review student proposals and manually compare them with past submissions to determine if the idea is original or similar to existing ones. This process is labor- intensive and time-consuming. With this system, however, the entire process is automated. Students will submit their project details, including the **title, abstract, and category**. Once submitted, the system will generate **sentence embeddings** for the title and abstract using **pre-trained Sentence Transformers models**: **paraphrase-MiniLM-L6-v2**, **paraphrase-multilingual-MiniLM-L12-v2**, **paraphrase-MPNet-base-v2**, and **All-MPNet- base-v2** [3]. These models convert the input text into **vector representations**, allowing for accurate similarity measurement through **cosine similarity** [1]. A predefined **threshold** will be set to determine whether the proposal is sufficiently unique compared to previously stored projects. If the similarity score exceeds the threshold, the proposal will be **automatically rejected** by the system. However, if the similarity score is below the threshold, the proposal will be **forwarded to an authorized person**, such as a Coordinator or HOD, for further review andfinal approval. This automated approach ensures that only original ideas progress to the review stage while eliminating duplicate or repetitive proposals early in the process.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. RELATED WORK

1. All-MPNet-base-v2

Model:

The All-MPNet-base-v2 model is a general-purpose transformer designed to provide high-quality sentence embeddings across a variety of tasks. It combines the strengths of MPNet with optimizations for diverse semantic similarity challenges. This model excels in producing embeddings that generalize well across multiple datasets and tasks, making it highly versatile.

Performance:

Like the other models, this one also achieved **100% accuracy** in the survey in detecting similarities for the project proposals which are copied from the dataset. The model achieved **96% accuracy** for the slightly twisted abstract and it achieved **30% accuracy** for the projects which are not in the dataset(new) . The model performs least as compared to the other three models listed for the entirely new abstract.

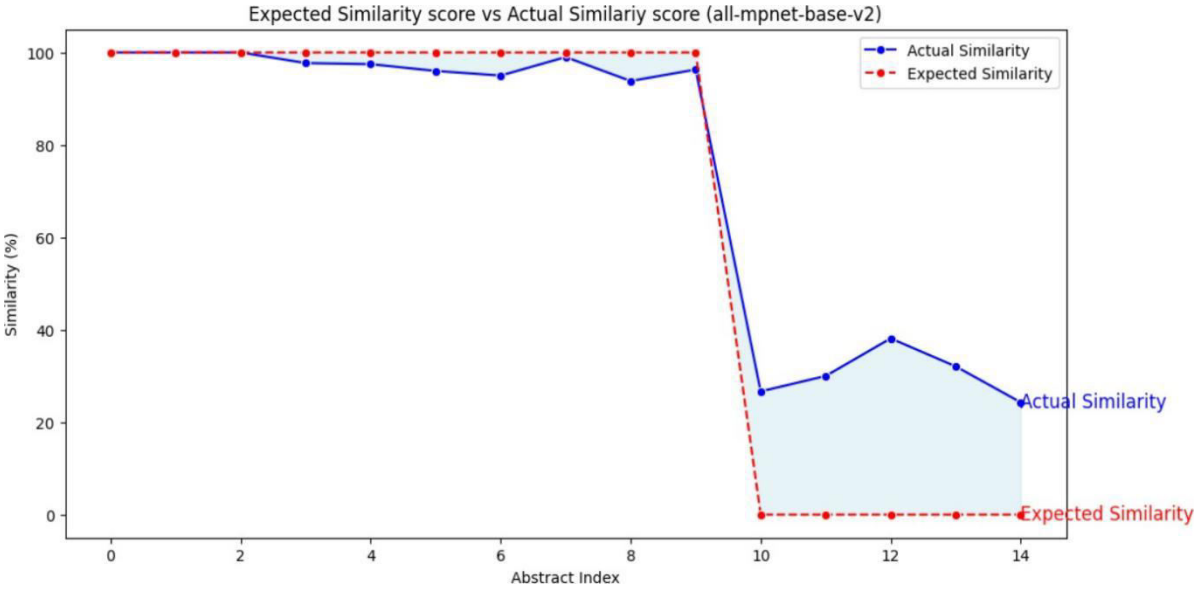

Figure-1: Performance of All-MpNet-Base-V2

Performance Metrics Summary (All-MpNet-Base-V2)

Metric	Value
Mean Absolute Error (MAE)	11.724666666666668
Mean Squared Error (MSE)	319.28004666666664
Root Mean Squared Error (RMSE)	17.86840918119648
Overall Accuracy (%)	98.35266666666666

2. paraphrase-MiniLM-L6-v2:

Model:

The paraphrase-MiniLM-L6-v2 is a compact transformer model designed for sentence embedding tasks [4]. It balances efficiency with accuracy, making it well-suited for real-time applications where low latency is required. Despite its smaller size, it provides strong performance in paraphrase detection tasks, making it ideal for similarity analysis.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Performance:

According to the survey data, this model achieved **100% accuracy** in detecting similarities for the project proposals which are copied from the dataset. The model achieved **90% accuracy** for the slightly twisted abstract and it achieved **47% accuracy** for the projects which are not in the dataset(new) . Its impressive performance indicates that it effectively identifies duplicate or highly similar abstracts, ensuring that repetitive project ideas are filtered out early in the review process.

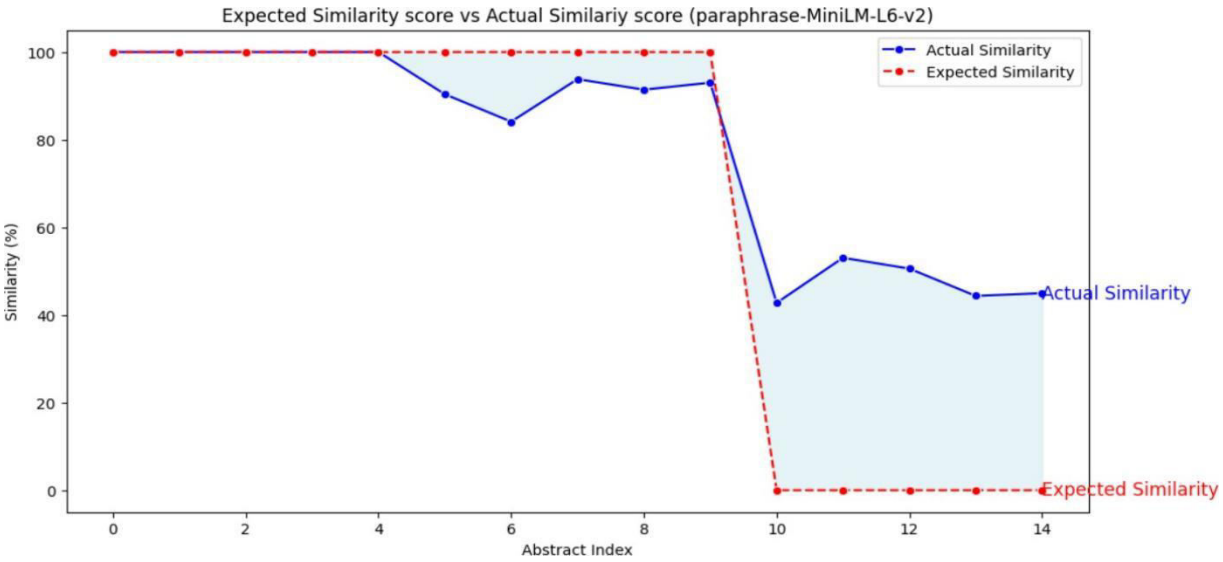

Figure-2: Performance of Paraphrase-MiniLM-L6-V2

Performance Metrics Summary (Paraphrase-MiniLM-L6-V2)

Metric	Value
Mean Absolute Error (MAE)	18.866666666666667
Mean Squared Error (MSE)	779.51372
Root Mean Squared Error (RMSE)	27.91977292171267
Overall Accuracy (%)	96.844

3. paraphrase-multilingual-MiniLM-L12-v2:

Model:

The paraphrase-multilingual-MiniLM-L12-v2 is an extension of the MiniLM series, optimized for multilingual tasks. This model is capable of generating embeddings across multiple languages, making it valuable for use cases involving diverse linguistic inputs [5]. It maintains the speed and lightweight architecture of the original MiniLM while expanding its applicability to multilingual datasets.

Performance:

The survey shows that this model also achieved **100% accuracy** in detecting similarities for the project proposals which are copied from the dataset. The model achieved **93% accuracy** for the slightly twisted abstract and it achieved **36% accuracy** for the projects which are not in the dataset(new) . Its strong results highlight its potential in ensuring that duplicate proposals are correctly flagged, even when dealing with submissions in different languages or varying linguistic structures. But the model slightly struggles with fully new data where its accuracy compared to **paraphrase-multilingual-MiniLM-L12-v2** is slightly less.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

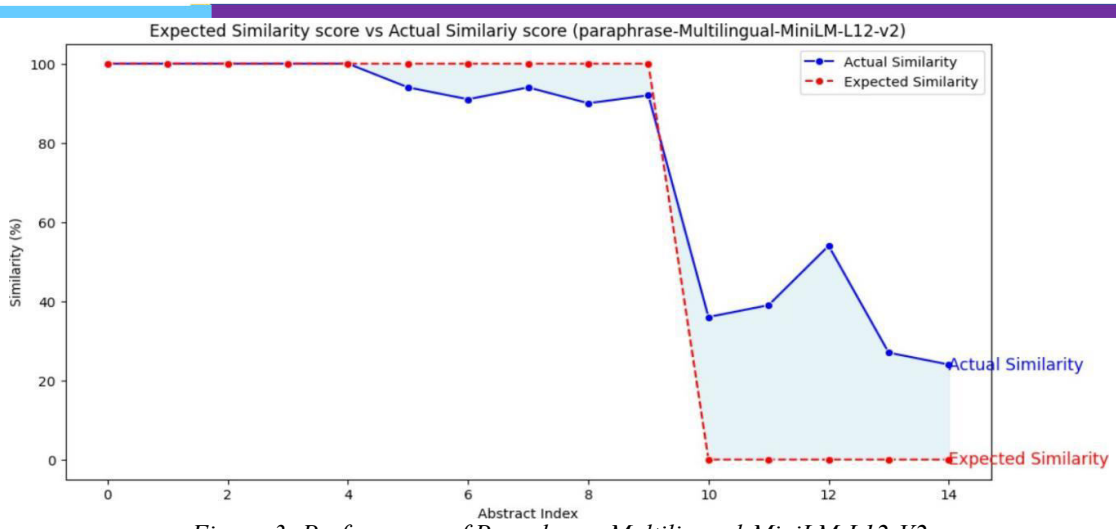

Figure-3: Performance of Paraphrase-Multilingual-MiniLM-L12-V2

Performance Metrics Summary (Paraphrase-Multilingual-MiniLM-L12-V2)

Metric	Value
Mean Absolute Error (MAE)	14.6
Mean Squared Error (MSE)	490.3333333333333
Root Mean Squared Error (RMSE)	22.143471573656495
Overall Accuracy (%)	97.4

4. paraphrase-MPNet-base-v2:

Model:

The paraphrase-MPNet-base-v2 model leverages MPNet, a masked language model known for capturing complex relationships between words in a sentence. It provides high-quality sentence embeddings, which are useful for tasks such as semantic similarity and paraphrase detection [6] [7]. MPNet models are particularly effective in scenarios that require deeper contextual understanding [8].

Performance:

Based on the survey data, this model performed with **100% accuracy** in detecting similarities for the project proposals which are copied from the dataset. The model achieved **95% accuracy** for the slightly twisted abstract and it achieved **39% accuracy** for the projects which are not in the dataset(new) , ensuring that project proposals with similar content are accurately identified [9]. Its strong performance underscores its ability to detect nuanced similarities in textual inputs.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

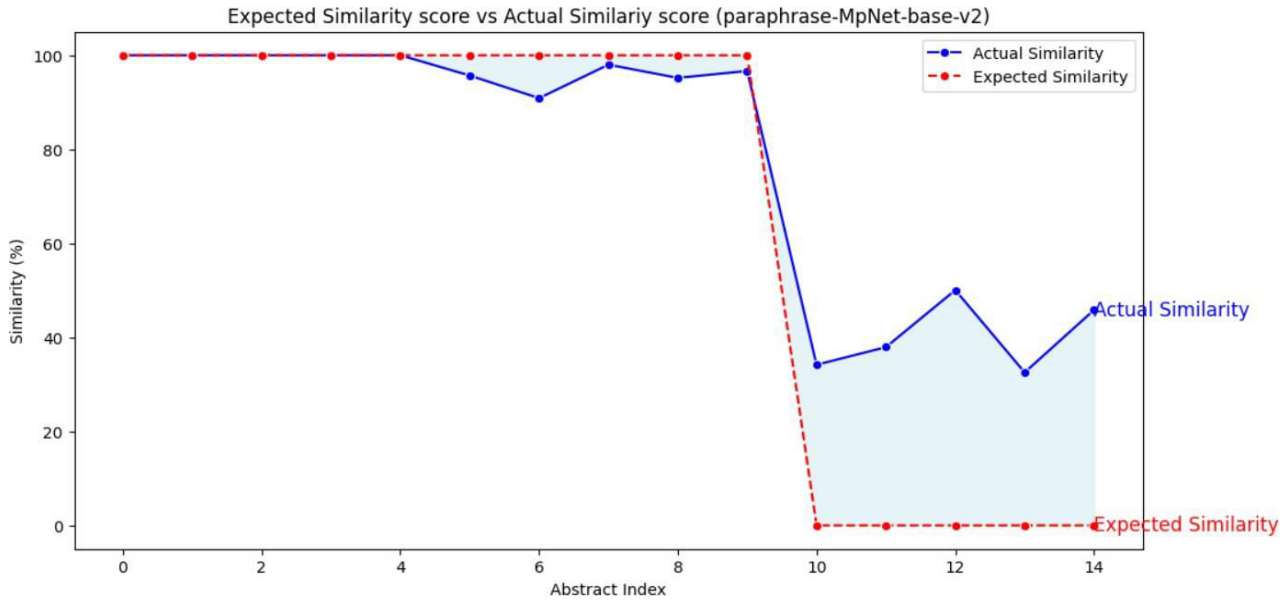

Figure-4: Performance of Paraphrase-MpNet-Base-V2

Performance Metrics Summary (Paraphrase-MpNet-Base-V2)

Metric	Value
Mean Absolute Error (MAE)	14.939333333333336
Mean Squared Error (MSE)	560.6533666666667
Root Mean Squared Error (RMSE)	23.678119998569706
Overall Accuracy (%)	98.42800000000001

III. PERFORMANCE COMPARISON TABLE

Abstracts	Prediction			
	paraphrase-MiniLM-L6-v2	All-MPNET-V2	paraphrase-multilingual-MiniLM-L12-v2	paraphrase-MPNet-base-v2
A fully differential calculation in perturbative quantum chromodynamicsis presented for the production of massive photon pairs at hadron colliders.	100%	100%	100%	100%
We present a critical review about the study of linear perturbations of matched spacetimes including gauge problems.	100%	100%	100%	100%

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This work presents a comprehensive differential calculation within the framework of perturbative quantum chromodynamics (QCD) for the generation of massive photon pairs at hadron colliders.	90.36%	96%	94%	95.69%

We investigate activated dynamics in a glassy system subjected to steady shear deformation through extensive numerical simulations.	93.0%	96.3%	92%	96.66%
This project investigates the effects of urban pollution on the biodiversity of local pollinator populations in urbangreen spaces.	42.75%	26.66%	36%	34.17%
This project explores the impact of varying light wavelengths on the photosynthetic efficiency of differentaquatic plants.	53.04%	30%	39%	37.92%

IV. RESULTS

while all models have their strengths, the All-MPNet-base-v2 strikes the best balance between accuracy on twisted abstracts and performance with new projects, making it the most robust choice for an automated similarity detection system in academic project proposals.

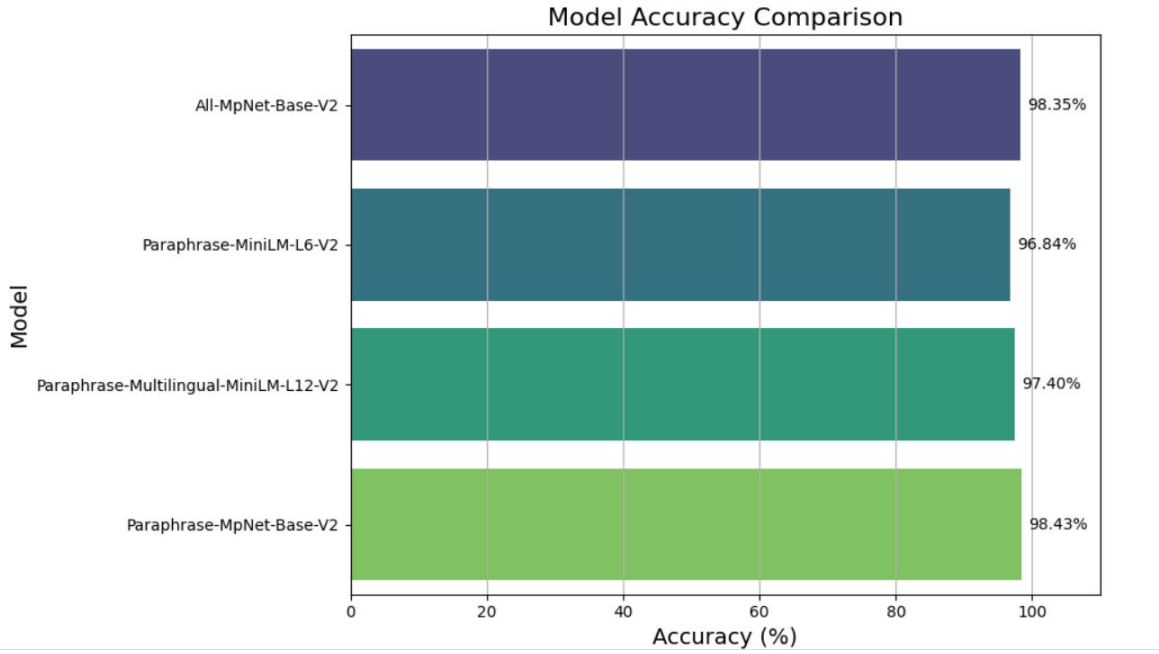

Figure-5: Model Accuracy Comparison

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. CONCLUSION

The system offers a significant improvement for academic institutions by **automating the manual task** of proposal evaluation. It ensures that faculty members can focus on more critical tasks while promoting **originality** in student projects. This solution not only streamlines the workflow but also ensures a fair and efficient review process. With future potential applications in other fields, such as **patent applications**, the system demonstrates a versatile approach to handling repetitive evaluation tasks through **AI-driven automation**.

REFERENCES

1. G. N. Reimers, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP),vol. 2, pp. 3973- 3983, 2019.
2. Mikolov, T., Chen, K., Corrado, G., & Dean, J., "Efficient estimation of word representations in vector space," Proceedings of the International Conference on Learning Representations (ICLR), vol. 2, pp. 1-12, 2013.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, vol. 1, pp. 4171-4186., 2019.
4. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," Proceedings of the International Conference on LearningRepresentations (ICLR), vol. 8, pp. 1-11, 2020.
5. Schuster, M., & Nakajima, K., "Japanese and Korean voice search," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 5149-515, 2012.
6. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., "Deep contextualized word representations," Proceedings of NAACL-HLT, vol. 2, pp. 2227-2237, 2018.
7. Pennington, J., Socher, R., & Manning, C. D., "GloVe: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), vol. 14, pp. 1532-1543, 2014.
8. Johnson, R., Zhang, T., & LeCun, Y., "Deep pyramid convolutional neural networks for text categorization," Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL), vol. 2, pp. 562-570, 2015.
9. Wang, A., Singh, A., Michael, J., Hill, F., Levy, , "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," Proceedings of ICLR 2019, vol. 3, pp. 1-12, 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details