IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

VOLUME : 11 ISSUE : 02 Print / Issue Publication Date: June 2026

ISSN : 2455-2143

DOI : 10.33564/IJEAST.2026.v11i02.009

Indexed In

Google SCHOLAR · OPEN ACCESS · doi · Crossref

WWW.IJEAST.COM

editor@ijeast.com

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 02, ISSN No. 2455-2143, Pages 81-86
Published Online June 2026 in IJEAST (http://www.ijeast.com)

DEEP-MATCH: A TRANSFORMER-BASED SEMANTIC RESUME SCREENING FRAMEWORK USING SENTENCE-BERT AND COSINE SIMILARITY

Anjali Mishra, Prof. Amisha Patodi
Department of Computer Science and Engineering,
Sanghvi Institute of Management and Science (SIMS),
Rau, Indore, Madhya Pradesh, India
(Affiliated to RGPV, Bhopal)

Abstract— **Modern organizations face an escalating bottleneck in talent acquisition. The exponential growth of online job applications has overwhelmed the capacity of human-driven screening processes. Conventional Applicant Tracking Systems, built on term-frequency heuristics and string-matching logic, fail to bridge the semantic gap between how candidates describe their competencies and how job descriptions articulate requirements. This paper introduces Deep-Match, a modular, production-ready recruitment intelligence framework that resolves this semantic disconnect through the deployment of Sentence-BERT (SBERT), specifically the all-MiniLM-L6-v2 architecture, to generate context-aware 384-dimensional document embeddings. Candidate resumes are ingested via a coordinate-aware PDF parser built with PyMuPDF, normalized through a multi-stage preprocessing pipeline, and vectorized using a pre-trained transformer encoder with mean pooling. Job descriptions undergo an identical encoding pathway, after which Cosine Similarity serves as the ranking metric. A threshold-driven classification engine categorizes candidates into Excellent, Average, or Low match tiers. The system further incorporates a Skill Gap Analysis module that surfaces missing competencies relative to the target role. Experimental evaluation demonstrates that Deep-Match reduces initial screening time by over 75% while achieving substantially higher ranking precision than TF-IDF and Word2Vec baselines. The complete stack is implemented using FastAPI, React, MySQL, and ChromaDB, containerized with Docker, and operates entirely on commodity CPU hardware.**

Keywords— **Natural Language Processing, Sentence-BERT, Transformer Architecture, Cosine Similarity, Resume Screening, Applicant Tracking System, Semantic Matching, HR Automation, FastAPI, ChromaDB.**

I. INTRODUCTION

The digital transformation of the global employment market has produced a compounding challenge for Human Resource departments worldwide. Online job platforms have dramatically lowered the barrier to applying for positions, resulting in organizations routinely receiving hundreds of applications for a single vacancy. Research from industry analysts shows that recruiters spend an average of six seconds on an initial resume review, creating an environment where qualified talent is systematically overlooked not due to lack of merit but due to the sheer volume of applications.

The predominant technological response has been the Applicant Tracking System (ATS). However, most commercial ATS products remain anchored in lexical heuristics: a candidate whose resume contains the specific string matching those in the job description is ranked higher, irrespective of professional context or semantic equivalence of alternative phrasings.

This lexical paradigm introduces a critical failure mode. A candidate describing themselves as a "Machine Learning Engineer" may be excluded from a search targeting a "Data Scientist" role, despite possessing an 85-90% overlap in actual technical competency. This vocabulary mismatch problem is endemic to a domain where roles and tools evolve faster than standardized taxonomies can track.

This research proposes Deep-Match, which reorients resume screening from lexical matching to semantic understanding. By encoding resumes and job descriptions into a shared high-dimensional semantic space using SBERT, the framework enables geometric distance metrics to quantify candidate suitability. The end-to-end processing pipeline is shown in Fig. 1.

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 02, ISSN No. 2455-2143, Pages 81-86
Published Online June 2026 in IJEAST (http://www.ijeast.com)

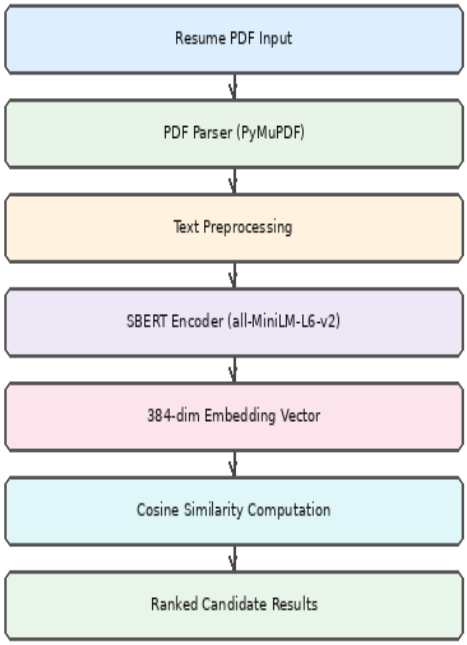

Fig. 1. Deep-Match End-to-End Processing Pipeline

The contributions of this work are threefold. First, it demonstrates the practical application of transformer-based semantic embeddings to professional document matching. Second, it presents a complete production-grade system integrating modern backend, frontend, and vector database technologies within a containerized deployment. Third, it introduces a Skill Gap Analysis module providing interpretable output beyond a simple numerical score.

II. RELATED WORK

A. Statistical Foundations of Document Retrieval

The study of automated document similarity is rooted in the Information Retrieval discipline. The TF-IDF weighting scheme formalized by Salton and Buckley in 1988 provided the first mathematically principled approach to term importance within a corpus [7]. While computationally efficient, its Bag-of-Words representation discards sequential information entirely. Latent Semantic Analysis introduced by Deerwester et al. in 1990 attempted to recover latent semantic structure through Singular Value Decomposition [6]. However, LSA is expensive to scale and cannot resolve polysemy.

B. Neural Word Embeddings

Word2Vec by Mikolov et al. in 2013 produced dense vector representations encoding semantic relationships through spatial proximity [4]. GloVe by Pennington et al. extended this by incorporating global co-occurrence statistics [5]. Despite their contributions, both produce static embeddings where a token's representation is fixed regardless of context, rendering them unsuitable for polysemous terms central to professional document analysis.

C. Transformer Architectures and SBERT

The Transformer architecture introduced by Vaswani et al. in 2017 enabled a fundamentally new approach through self-attention [3]. BERT by Devlin et al. in 2018 demonstrated state-of-the-art performance through bidirectional context encoding [1]. Reimers and Gurevych identified that standard BERT requires cross-encoder architecture for similarity tasks, resulting in quadratic complexity. Sentence-BERT resolves this through a Siamese network enabling independent pre-computation of sentence embeddings [2], as illustrated in Fig. 2.

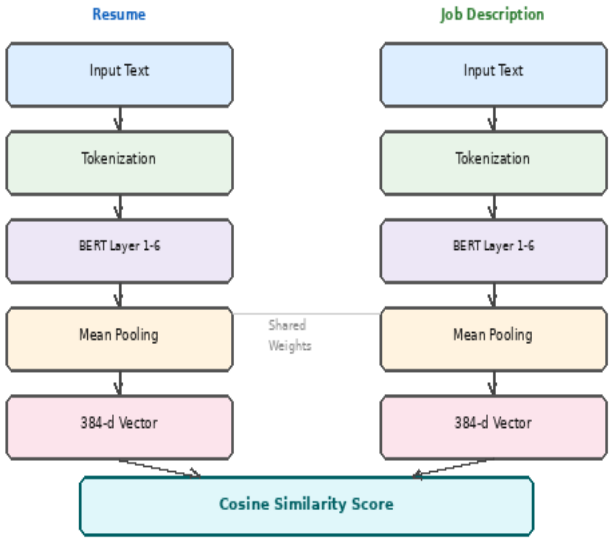

Fig. 2. Sentence-BERT Siamese Network Architecture

III. PROPOSED METHODOLOGY

A. System Architecture Overview

The Deep-Match framework is designed according to modular microservice architecture principles. The system comprises four primary computational layers: Data Ingestion, Preprocessing and Tokenization, Neural Embedding Engine, and Similarity Analytics, supported by MySQL for structured data and ChromaDB for vector storage. The complete technology stack is shown in Fig. 3.

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 02, ISSN No. 2455-2143, Pages 81-86
Published Online June 2026 in IJEAST (http://www.ijeast.com)

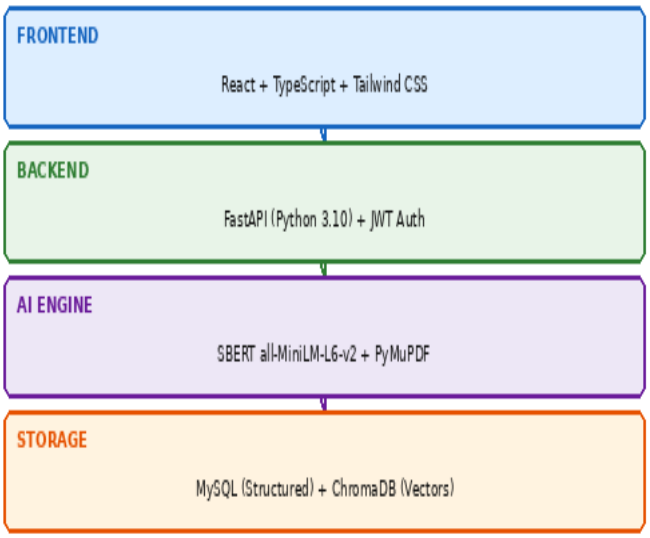

Fig. 3. Deep-Match System Architecture and Technology Stack

B. Data Ingestion and Document Parsing

Professional resumes present a significant parsing challenge due to the diversity of structural formats. Naive sequential PDF text extraction often results in scrambled text where sidebar content is interleaved with body paragraphs. Deep-Match addresses this through coordinate-based extraction using PyMuPDF. The parser retrieves all text blocks with bounding box coordinates, sorts blocks by vertical then horizontal position, and reconstructs logical reading order regardless of the original layout.

C. Preprocessing Pipeline

The normalized text undergoes a multi-stage preprocessing pipeline: lowercase normalization to eliminate case-based token duplication, regex-based sanitization to remove web artifacts, and sentence segmentation. WordPiece tokenization handles out-of-vocabulary terms by decomposing unknown words into constituent subword units, which is particularly valuable for technical resume content where version-specific tool names appear with high frequency.

D. Sentence-BERT Embedding Engine

The core module loads the all-MiniLM-L6-v2 SBERT model at application startup. Input tokens are processed through six transformer encoder layers applying multi-head self-attention followed by position-wise feed-forward transformations. Output token embeddings are aggregated through mean pooling to produce a single 384-dimensional document-level vector. Formally, given document D tokenized into n tokens:

$$E(D) = (1/n)\ \Sigma\ BERT_6(token_i),\ \ i = 1 \text{ to } n$$

E. Similarity Scoring and Classification

Candidate ranking is performed through Cosine Similarity between the job description embedding Vjd and each candidate resume embedding Vr:

$$Similarity(Vjd, Vr) = (Vjd \cdot Vr) / (||Vjd|| \times ||Vr||)$$

The classification thresholds are illustrated in Fig. 5: scores at or above 0.80 are Excellent Match, 0.50-0.79 are Average Match, and below 0.50 are Low Match, determined through empirical testing on a manually annotated validation set.

F. Skill Gap Analysis Module

The Skill Gap Analysis module performs token-level intersection analysis between technical skills extracted from the job description and those present in the candidate resume, as illustrated in Fig. 4. Skills identified in the job description but absent from the resume are surfaced as a structured list of missing competencies, providing interpretable, auditable information about the nature of the gap.

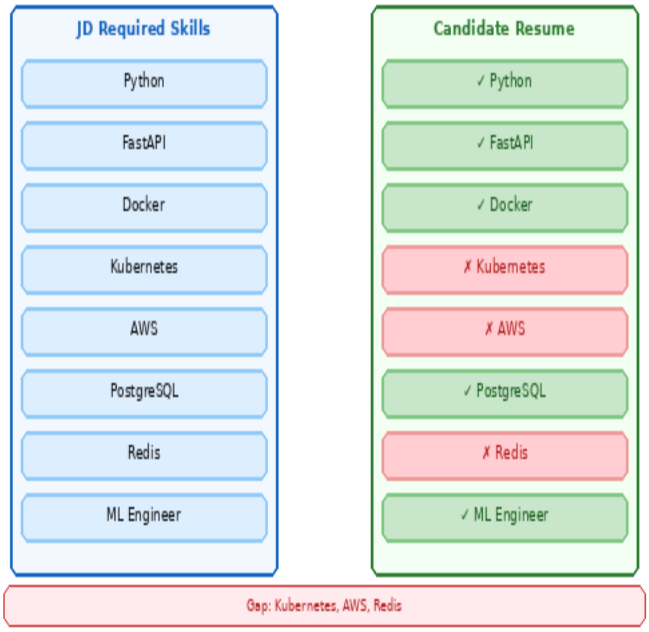

Fig. 4. Skill Gap Analysis — JD Requirements vs Candidate Resume

IV. SYSTEM IMPLEMENTATION

A. Backend API Architecture

The backend is implemented as a FastAPI application with four primary route modules: authentication, job management, candidate management, and the matching engine. The authentication module implements JWT-based

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 11, Issue 02, ISSN No. 2455-2143, Pages 81-86
Published Online June 2026 in IJEAST (http://www.ijeast.com)

stateless session management. The candidate upload endpoint accepts multipart form data, invokes the PDF parser, generates an SBERT embedding, and persists it to ChromaDB. The structured candidate record is simultaneously written to MySQL, creating a consistent reference between the relational and vector stores.

B. Vector Database Integration
ChromaDB serves as the vector storage and retrieval layer, maintaining resume embeddings with cosine distance as the configured metric. The embedded mode deployment runs directly within the backend process, eliminating network overhead and reducing operational complexity. Pre-computed embeddings enable sub-millisecond retrieval for returning candidates in subsequent matching requests.

C. Frontend Implementation
The React and TypeScript frontend provides four primary views: a Dashboard with aggregate recruitment statistics and chart visualization; a Jobs view for creating and managing job descriptions; a Candidates view with drag-and-drop PDF upload and real-time processing feedback; and the AI Match view enabling one-click execution of the full semantic matching pipeline, presenting ranked results with score gauges, classification badges, and expandable skill gap details.

D. Experimental Setup
Development and evaluation were conducted on a standard workstation with an Intel Core i5 processor and 8GB RAM without a dedicated GPU. All transformer inference runs on CPU using the PyTorch backend, demonstrating that production-viable semantic matching does not require specialized hardware. The complete stack is containerized using Docker Compose enabling single-command deployment across operating systems.

V. RESULTS AND EVALUATION

A. Comparative Performance Analysis
The Deep-Match framework was evaluated against TF-IDF with cosine similarity and Word2Vec with averaged word vectors. A test set of 50 resume-job description pairs was constructed with ground-truth relevance labels assigned through manual expert evaluation. Table I and Fig. 6 present the results across Precision@5 and Mean Reciprocal Rank (MRR) metrics.

Method	Precision@5	MRR	Time Reduction
TF-IDF + Cosine	0.54	0.61	48%
Word2Vec (Avg)	0.62	0.69	58%
Deep-Match (SBERT)	**0.84**	**0.89**	**76%**

Table I. Comparative Evaluation of Screening Methods

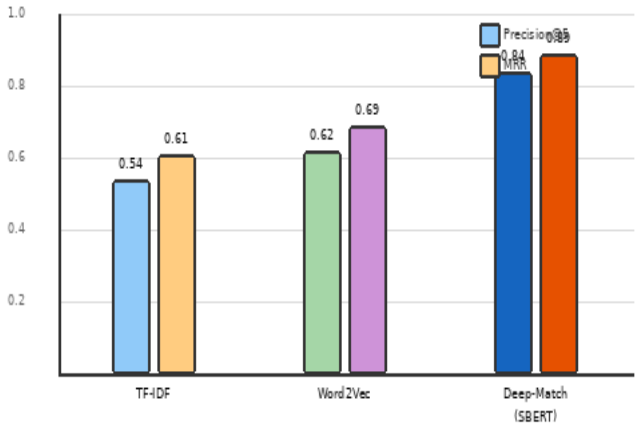

Fig. 6. Performance Comparison: Precision@5 and MRR Across Methods

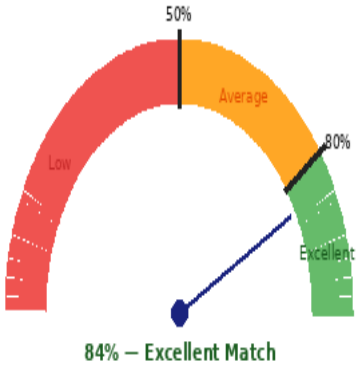

Fig. 5. Match Score Classification Gauge — Candidate Tiering Thresholds

Deep-Match achieves Precision@5 of 0.84, a 55.6% improvement over TF-IDF and 35.5% over Word2Vec. The

MRR of 0.89 confirms the first relevant candidate appears near the top of the ranked list in the majority of test cases.

B. Semantic Gap Resolution

Qualitative analysis confirms the framework successfully resolves vocabulary mismatches that defeat lexical matching. A job description requiring a "Backend Developer proficient in RESTful services" correctly ranked a resume describing "API development using Django REST Framework" in the top tier, despite sharing no exact keyword overlap. Similarly, Quantitative Analyst resumes are correctly matched to Data Scientist job descriptions based on shared statistical and computational concepts in the semantic vector space.

C. Computational Performance

SBERT encoding of a single document requires approximately 35-80 milliseconds on CPU depending on document length. For a batch of 50 candidate resumes against one job description, the complete matching pipeline completes in under 8 seconds without GPU acceleration, confirming practical viability for SME deployment scenarios.

VI. DISCUSSION

A. Advantages Over Lexical Methods

The performance differential validates the central thesis: semantic matching through contextual transformer embeddings captures professional relevance that character-level methods cannot. The self-attention mechanism enables the model to weight terminology differently based on surrounding context. A candidate who mentions Python in the context of "scripting automation pipelines" receives a different embedding contribution than one who mentions it in "introductory exposure," despite identical surface-level keyword presence.

B. Interpretability and Bias Reduction

The Skill Gap Analysis module addresses the opacity concern by providing an interpretable, auditable explanation for each match score. Recruiters can verify that the system reasoning aligns with their professional judgment. Furthermore, because ranking is derived from professional content semantics rather than demographic signals, the system promotes a skills-first evaluation paradigm reducing the influence of implicit human bias on initial screening decisions.

C. Limitations and Future Work

The current Skill Gap module operates on a fixed keyword vocabulary. Future work will replace this with attention-weight-based extraction identifying semantically important spans directly from transformer internal representations. The classification thresholds were determined on a moderate validation set; a larger-scale study across diverse industry sectors would enable data-driven threshold calibration. Integration of structured candidate data beyond resume text represents a natural extension pathway.

VII. CONCLUSION

This paper has presented Deep-Match, a production-ready transformer-based semantic resume screening framework that demonstrably outperforms conventional lexical matching approaches. By leveraging Sentence-BERT to encode professional documents into a shared semantic vector space and employing Cosine Similarity as the ranking metric, the system resolves the vocabulary mismatch problem that has long undermined automated ATS tools. The framework operates without GPU hardware, is fully containerized, provides interpretable skill gap output, and achieves a 76% reduction in recruiter screening effort. Deep-Match represents a meaningful contribution to applied AI literature on intelligent HR automation, and its modular architecture provides a foundation for continued enhancement through domain-adaptive fine-tuning and expanded interpretability mechanisms.

VIII. REFERENCES

[1]. Devlin J., Chang M., Lee K., and Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, Minneapolis, MN, USA, (pp. 4171-4186). DOI: 10.18653/v1/N19-1423.

[2]. Reimers N., and Gurevych I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP-IJCNLP, Hong Kong, China, (pp. 3982-3992). DOI: 10.18653/v1/D19-1410.

[3]. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., and Polosukhin I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, vol. 30, (pp. 5998-6008).
DOI: 10.48550/arXiv.1706.03762.

[4]. Mikolov T., Sutskever I., Chen K., Corrado G.S., and Dean J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems, vol. 26, (pp. 3111-3119).
DOI: 10.48550/arXiv.1310.4546.

[5]. Pennington J., Socher R., and Manning C.D. (2014). GloVe: Global Vectors for Word Representation. Proceedings of EMNLP, Doha, Qatar, (pp. 1532-1543). DOI: 10.3115/v1/D14-1162.

[6]. Deerwester S.C., Dumais S.T., Landauer T.K., Furnas G.W., and Harshman R.A. (1990). Indexing by Latent Semantic Analysis. Journal of the

American Society for Information Science, vol. 41, no. 6, (pp. 391-407). DOI: 10.1002/asi.4630410602.

[7]. Salton G., and Buckley C. (1988). Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management, vol. 24, no. 5, (pp. 513-523). DOI: 10.1016/0306-4573(88)90021-0.

[8]. Navigli R. (2009). Word Sense Disambiguation: A Survey. ACM Computing Surveys, vol. 41, no. 2, (pp. 1-69). DOI: 10.1145/1459352.1459355.

[9]. Maheshwary S., and Pande S. (2018). Matching Resumes to Jobs via Deep Siamese Network. Proceedings of the ACL Workshop on Relevance of Linguistic Structure in Neural Architectures for NLP, Melbourne, Australia, (pp. 1-8). DOI: 10.18653/v1/W18-2901.

[10]. Goodfellow I., Bengio Y., and Courville A. (2016). Deep Learning. MIT Press, Cambridge, MA, USA, (pp. 326-366). ISBN: 978-0262035613.

[11]. Bird S., Klein E., and Loper E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Sebastopol, CA, USA, (pp. 1-479). ISBN: 978-0596516499.

[12]. He K., Zhang X., Ren S., and Sun J. (2016). Deep Residual Learning for Image Recognition. Proceedings of IEEE CVPR, Las Vegas, NV, USA, (pp. 770-778). DOI: 10.1109/CVPR.2016.90.

[13]. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., and Stoyanov V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint, (pp. 1-13). DOI: 10.48550/arXiv.1907.11692.

[14]. Peters M.E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., and Zettlemoyer L. (2018). Deep Contextualized Word Representations. Proceedings of NAACL, New Orleans, LA, USA, (pp. 2227-2237). DOI: 10.18653/v1/N18-1202.

[15]. Luo X., An B., Jiang Y., and Bing L. (2019). ResumeGAN: Towards Realistic Resume Generation via Representation Learning. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, (pp. 1897-1906). DOI: 10.1145/3357384.3357873.

IJEAST

INTERNATIONAL JOURNAL

OF ENGINEERING APPLIED SCIENCE AND TECHNOLOGY

ABOUT IJEAST

International journal of Engineering Applied Science and Technology (IJEAST) is a peer-reviewed, open access journal that publishes high - quality research paper in the field of Engineering, Applied Science and Technology.

IJEAST aims to provide a platform for researchers, academicians, and professionals to share their innovative ideas, research findings, and practical experiences with the global scientific community.

FOCUS AREAS

- Engineering
- Applied Science
- Technology
- Innovation & Development
- Interdisciplinary Studies

PEER REVIEWED

All submission are rigorously peer reviewed to ensure quality.

OPEN ACCESS

Free and unrestricted access to research for all.

GLOBAL REACH

Connecting researchers and Professionals worldwide.

TIMELY PUBLICATION

We Ensure a swift and efficient publication process.

For more information, visit our website

www.ijeast.com

editor@ijeast.com

www.ijeast.com

India

2455-2143