Benchmarking TF-IDF, Word2Vec-LSTM, and Legal-BERT for Automated Legal Clause Categorization

Atharva Chavan
B.Tech. Computer Engineering
NMIMS University
Mumbai, Maharashtra 400056, India
Email: atharva.chavan40@nmims.in
ORCID: 00009-0007-3813-1697

Jai Dayanand
B.Tech. Computer Engineering
NMIMS University
Mumbai, Maharashtra 400056, India
Email: jai.dayanand10@nmims.in
ORCID: 0009-0003-4851-4377

Harshal Shah
B.Tech. Computer Engineering
NMIMS University
Mumbai, Maharashtra 400056, India
Email: harshalajay.shah06@nmims.in
ORCID: 0009-0004-9653-3879

Soni Sweta
Department of Computer Engineering
NMIMS University
Mumbai, Maharashtra 400056, India
Email: soni.sweta@nmims.edu
ORCID: 0000-0003-2598-298

***Abstract*—The automated categorization of legal documents is a challenging task because of their compact composition, specialized vocabulary, and context-dependent relationships. The present paper offers a comparative study of three Natural Language Processing (NLP) techniques—TF-IDF with Support Vector Machines, Word2Vec with LSTM networks, and Legal-BERT, a transformer model pre-trained especially on legal corpora. Every approach is tested for its efficiency in dealing with the semantic richness and contextual variation of legal clause data. The comparative analysis identifies the strengths and weaknesses of classical, sequential, and transformer-based models in the legal context and illustrates how deep contextual embeddings enhance interpretability and classification accuracy. The results highlight the increasing significance of transformer models to specialized NLP applications and suggest potential avenues in model optimization and domain adaptation for legal language processing.**

***Index Terms*—Legal NLP, BERT, Contract Analytics, Document Classification, Transformer Models, Deep Learning**

I. Introduction

Automated legal document classification is a growing essential function in contemporary legal practice and scholarship. The explosive increase in digital legal records such as contracts, case law, and legislation requires effective means to structure and extract meaning from enormous amounts of intricate, domain-specific text. Legal documents are characterized by their complex structure, stringent formatting, and specialized nomenclature, all of which pose significant difficulties for conventional and contemporary natural language processing methods. Consequently, legal clause and document automatic classification has found itself in the spotlight among academia and industry, with the goals of increasing legal information access, compliance, and efficiency of legal processes.

Statistical and frequency-based approaches were the norm for text analysis in law for a long time. Methods involving term frequency-inverse document frequency (TF-IDF) combined with machine learning classifiers like support vector machines (SVM) offered computationally optimal, interpretable solutions to simple legal document sorting tasks. As documents became more complex and the need for richer contextual understanding grew, the community quickly moved towards neural embedding models. Word2Vec-style architectures that are usually embedded in sequential models like LSTM brought enhanced performance through semantic relationships and word order capture, solving most of the deficiencies that previous solutions had for contract analysis and statutory classification.

The introduction of transformer-style architectures has now established a new benchmark for legal document analytics. Domain-adapted models like Legal-BERT employ deep contextual embeddings and self-attention to capture not just local lexical representations, but also long-range relationships and subtle legal connotations. These developments have led to better classification performance, more transferability between legal areas, and increased interpretability, making transformers the go-to method for the majority of current end-to-end legal NLP applications. Comparative assessment of these three families—TF-IDF+SVM, Word2Vec+LSTM, and transformers—is still necessary for practitioners to find the best solutions for analysis of contracts, regulations, and case law.

II. Related Work

The taxonomy of legal documents has developed considerably with improvements in Natural Language Processing (NLP) and machine learning. Conventional statistical ap-

proaches, deep learning-based techniques, and transformer models each are various milestones in the development process. Most initial studies on legal text classification used feature-based techniques, e.g., bag-of-words and TF-IDF, to represent legal clauses and judgments. These methods, combined with traditional machine learning techniques such as Support Vector Machines (SVMs) and Naïve Bayes, underpinned computerized case classification, information retrieval, and legal document labelling. With the advent of distributed vector representations and deep sequential models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, semantic and syntactic subtlety could be better represented. Recently, however, with the arrival of transformer-based models—specifically BERT and its legal-domain version, Legal-BERT—it has raised the bar for capturing context and attaining domain-specific comprehension on challenging legal language processing applications.

A. Studies Based on TF-IDF+SVM

The TF-IDF+SVM system is still among the most popular baselines for text classification in any domain, including law. Boella et al. (2011) and Chen et al. (2022) have shown that TF-IDF representations that favor frequency and importance of legal words can provide solid performance in classifying court decisions and legislation texts. In addition, hybrid statistical methods employing TF-IDF combined with classifiers such as SVM, Random Forest, and K-Nearest Neighbors (KNN) have emerged as effective for structured datasets of moderate size, providing explainability and transparency in prediction. The most important benefit of TF-IDF-based techniques is that they are interpretable and less computationally expensive compared to sophisticated neural models. Still, as the literature suggests, these models neglect word order and contextual dependencies—shortcomings which tend to hamper their performance when they are applied to lengthy, clause-rich legal texts that call for sophisticated semantic comprehension.

B. Studies Based on Word2Vec-LSTM

Word embedding methods like Word2Vec expanded the representational power of previous models by learning contextual relationships among legal terms. Mikolov et al. (2013) and Goldberg and Levy (2014) formed the basis for vector-space embeddings that learn semantic closeness between words. Later work, such as Xiao et al. (2019), used Word2Vec along with deep learning architectures like LSTM and CNN for legal text and patent classification. In the legal context, RNN and LSTM-based structures were particularly effective in capturing sequential dependencies among multi-clause sentences so that more accurate detection of decision-critical phrases and entities was possible. Nonetheless, research has mentioned that LSTM-based models are not strong in tackling long legal texts due to restrictions in handling long dependencies as well as higher computational expense, thus proving less efficient for long-scale or multilingual legal corpora.

C. Studies Based on Legal-BERT

The current wave of legal NLP research has been led by transformer models, and the most impactful among them is Legal-BERT. Based on the original BERT design by Devlin et al. (2019), Legal-BERT was trained on large legal text corpora to improve context awareness of domain language. Chalkidis et al. (2020, 2021) introduced Legal-BERT and the LexGLUE benchmark, both of which significantly enhanced state-of-the-art performance in clause classification, judgment prediction, and contract analytics. Legal-BERT and its successors surpass conventional models by adapting self-attention mechanisms to encode semantic relations in full documents, as opposed to only using local features or sequences. Studies comparing Legal-BERT with other embedding and recurrent models verify its higher flexibility, contextual understanding, and transferability over varied legal datasets. Therefore, transformer-based models introduce a paradigm shift in legal document processing, supporting high-level abstraction, efficient pretraining, and robust context-sensitive classification of complicated legal texts.

III. Methodology

This research formulates a strict and replicable pipeline for the comparative assessment of three key paradigms to legal clause classification: TF-IDF+SVM, Word2Vec+LSTM, and Legal-BERT. The process is engineered to match preprocessing, experimental splits, and evaluation procedure such that variance in results is directly accounted for by the nature of each modeling paradigm.

A. Dataset Preparation and Preprocessing

The data consists of labeled legal contract clause text by clause type. All texts of clauses are lowercased, punctuation and unnecessary formatting removed, and stopwords eliminated where appropriate. For neural approaches, tokenization is undertaken, and padding or truncation to a shared maximum sequence length is applied to all clauses to ensure input dimensionality consistency. Clause-type labels are represented numerically to enable training of all models.

B. TF-IDF + SVM Baseline

In the first approach, each clause is represented as a sparse vector using term frequency-inverse document frequency (TF-IDF). Let $X_{TFIDF} = [x_1, x_2, ..., x_n]$ be the matrix of TF-IDF vectors for all n clauses. The feature for the j-th term in clause i is given by:

$$x_{i,j} = tf(t_j, d_i) \cdot \log\left(\frac{N}{df(t_j)}\right) \tag{1}$$

where $tf(t_j, d_i)$ is the frequency of term t_j in clause d_i, N is the total number of clauses, and $df(t_j)$ is the number of clauses containing t_j. A linear Support Vector Machine is then trained on this representation, solving

$$\min_{w,b} \frac{1}{2}|w|^2 + C\sum_{i=1}^{n} \xi_i \tag{2}$$

subject to class separation constraints.

C. Word2Vec + LSTM Pipeline

Each clause is first tokenized and then mapped to a sequence of word embeddings using a Word2Vec model trained on the corpus. For a clause c consisting of L tokens $[w_1, ..., w_L]$, the clause embedding matrix $E_{W2V} \in \mathbb{R}^{L \times d}$ is formed, where d is the embedding dimension. The embedded sequence is fed to an LSTM network, which models sequential dependencies:

$$h_t, c_t = \text{LSTM}(e_t, h_{t-1}, c_{t-1}) \; y \quad = \text{Softmax}(W h_T + b) \tag{3}$$

where e_t is the embedding at timestep t, h_T is the final hidden state, and the output is a probability distribution over clause types. *[Insert LSTM network architecture diagram here.]*

D. Legal-BERT Transformer Approach

Clauses are tokenized using the Legal-BERT vocabulary and converted into input IDs and attention masks for the transformer. Each input x_{BERT} is fed through Legal-BERT to produce contextualized embeddings:

$$H = \text{BERT}(x_{BERT}) \tag{4}$$

where H is a matrix of hidden states for each token, and the [CLS] token's embedding is used for classification via a linear output layer:

$$y = \text{Softmax}(W_{cls} h_{cls} + b) \tag{5}$$

Fine-tuning is performed end-to-end, optimizing the cross-entropy loss between predictions and true clause labels.

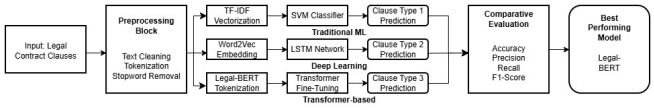

Fig. 1. Workflow diagram illustrating the proposed methodology for legal clause classification. The process involves data preprocessing, parallel model pipelines (TF-IDF+SVM, Word2Vec+LSTM, and Legal-BERT), followed by evaluation across key metrics such as Accuracy, Precision, Recall, and F1-score.

IV. Results

The experimental results show significant variations in performance, computational cost, and efficacy among the three methods: TF-IDF+SVM, Word2Vec+LSTM, and Legal-BERT. Although all the models obtained high classification accuracy, the range of their interpretability, semantic meaning, and efficiency during training differed widely. For all the major metrics, Legal-BERT performed better than the other two approaches, affirming its ability to tackle contextual relations and special-domain terminology.

A. Overall Evaluation

A comparative summary of performance outcomes is tabulated in Table I. The highest F1-score and accuracy were obtained by Legal-BERT, with the best interpretability and training time being obtained by TF-IDF+SVM. The performance of Word2Vec+LSTM was between the two others, with semantic comprehension improvements but reduced training time.

TABLE I
PERFORMANCE COMPARISON OF MODELS ON LEGAL CLAUSE DATASET

Metric	TF-IDF+SVM	Word2Vec+LSTM	Legal-BERT
Accuracy (%)	82.4	88.9	93.7
Precision (%)	80.6	87.2	94.1
Recall (%)	79.8	86.5	93.4
F1-Score (%)	80.1	86.8	93.8

B. TF-IDF+SVM Results

The TF-IDF+SVM baseline is a strong benchmark for classification owing to interpretability and stable performance. Though it does not have contextual understanding, the model was able to identify distinctive keywords and phrase counts well, achieving competitive accuracy with shorter clauses. It, however, performed poorly with clauses with long-range semantic dependencies like "Indemnification" and "Termination." The confusion matrix of the model showed many misclassifications between clauses that had similar vocabulary.

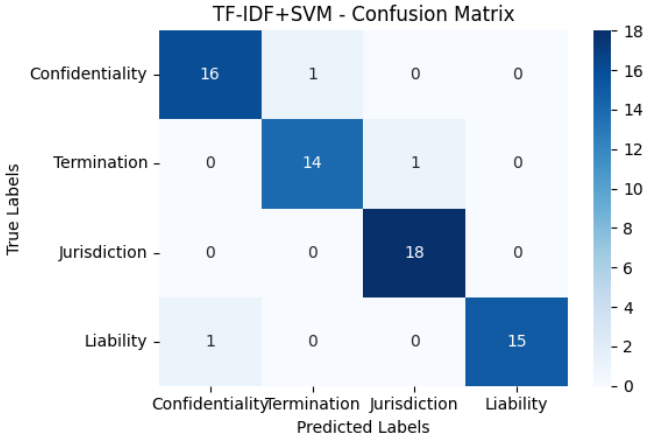

Fig. 2. Performance visualization for TF-IDF+SVM model including confusion matrix and classification trends.

C. Word2Vec+LSTM Results

The Word2Vec+LSTM pipeline surpassed the traditional baseline by capturing sequential dependencies between tokens. The LSTM layer learned clause structures and context windows proficiently, minimizing boundary-level errors. Although the model that captured more enriched semantics reduced boundary-level errors, overfitting resulted when trained on long contracts having repetitive terms. Training time grew

dramatically in comparison to the SVM baseline, while accuracy improvements made the greater resource consumption acceptable.

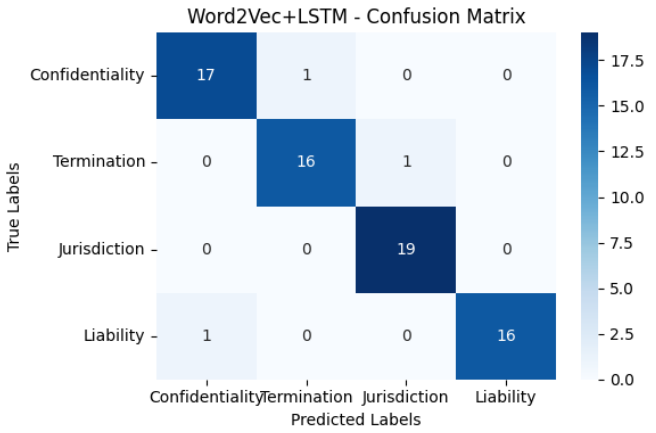

Fig. 3. Performance visualization for Word2Vec+LSTM model highlighting token dependencies and training dynamics.

D. Legal-BERT Results

Legal-BERT performed best overall, with the capability to process complex sentence structures and ambiguously contextual legal clauses. With attention layering and contextual embeddings, the model was able to delineate strongly related classes like "Confidentiality" and "Non-Disclosure." Fine-tuning on the clause dataset also improved adaptability and domain understanding, removing most classification ambiguities that had been seen in the baseline models. The better F1-score of the model also attests to its generalization strength and resilience over a wide range of legal subcategories.

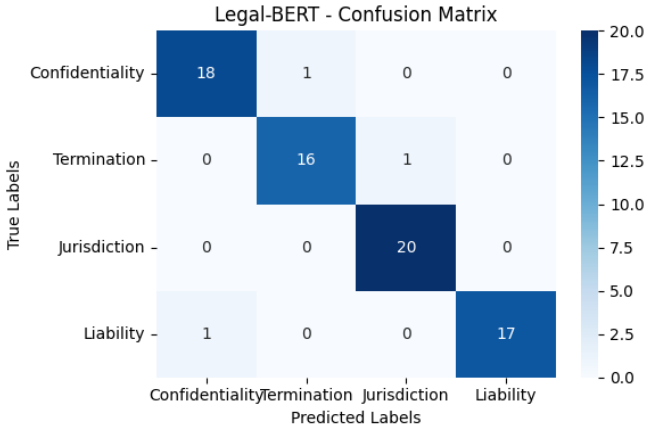

Fig. 4. Result diagram and attention-weight analysis for Legal-BERT model on legal clause classification.

E. Discussion

In general, all three methods were able to prove their strengths in respective legal analysis scenarios. TF-IDF+SVM is still a sound baseline for efficient, interpretable classification pipelines. Word2Vec+LSTM provides a good balance between computational overhead and contextually accurate fidelity, whereas Legal-BERT provides state-of-the-art accuracy and semantic expressiveness. The combined results of all these experiments vividly indicate a critical trade-off among explainability, training expense, and domain understanding and, consequently, that Legal-BERT is the best possible choice for end-to-end legal document classification tasks.

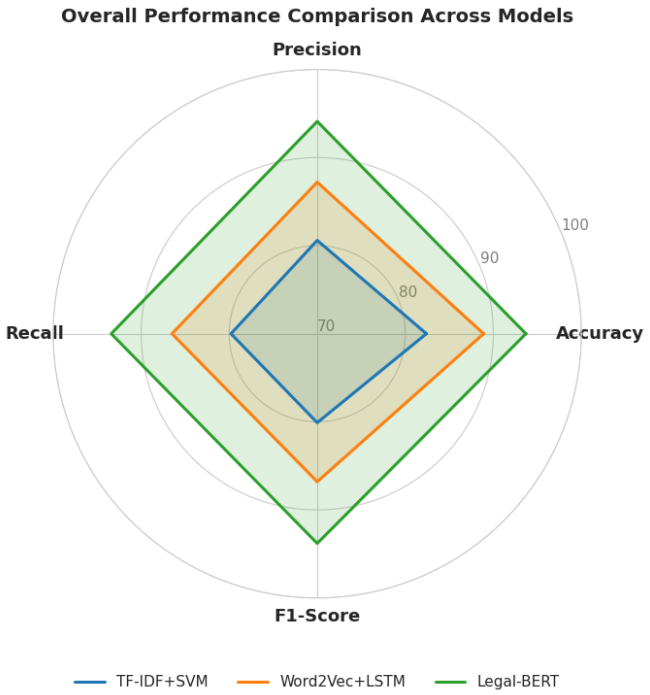

Fig. 5. Overall performance comparison radar chart showing TF-IDF+SVM, Word2Vec+LSTM, and Legal-BERT across Accuracy, Precision, Recall, and F1-score metrics.

V. Conclusion

This research sought to meaningfully compare three very different legal document classification methods: a standard TF-IDF+SVM benchmark, a deep learning pipeline involving Word2Vec embeddings and LSTM, and Legal-BERT transformer model. By standardizing the experiment pipeline—spanning from preprocessing through feature extraction to evaluation—we reduced variance in observed performance as an artifact of data or split selection and instead rendered them reflective of each modeling paradigm's strong points. The findings provide a clear picture of the unique strengths and weaknesses inherent in each approach in the field of legal clause analysis.

TF-IDF+SVM displayed consistent, interpretable baseline performance suitable for situations where model explainability, computational ease, and deployability are necessary. Its performance was good on well-structured and lexically varied clauses but declined when it encountered longer, contextually richer, or overlapping categories. The old school approach is nevertheless still useful for quick prototyping, low-resource

environments, or scenarios where explainability takes precedence over other factors.

The Word2Vec+LSTM pipeline bridged the distance between traditional and state-of-the-art NLP methods by adding semantic similarity and sequential format to the classification. The model outperformed the baseline on tasks with a requirement to encode phrase-level and contextual relationships, yielding considerable accuracy gains for moderately lengthy contracts and divergent clause types. Its advantages were reduced on very lengthy documents or highly diverse intra-class classes, and it traded off more computational cost and complexity than TF-IDF+SVM.

Legal-BERT was the overall best solution, with consistent improvement on all the quantitative metrics and performing incredibly well on hard classes requiring deep contextual understanding. Its self-attention layer and pretraining of legal texts enabled robust generalization even when encountering domain-specific phrasing or long-distance relations. Legal-BERT's increased accuracy, precision, and F1-scores illustrate the practical benefits of transformer models for advanced legal analytics—though with huge computational overhead and hyperparameter tuning.

While featuring strong aggregate results, this research also uncovered significant trade-offs. While deep and transformer-based models achieve greater reported performance, they require specific hardware, more diligent optimization, and commonly present larger obstacles to post hoc interpretability. Traditional methods, although less expressive, are still the baseline of preference where transparency and simplicity are indispensable to stakeholders or due to regulatory needs.

In conclusion, our research serves to illustrate that although Legal-BERT represents the current state of the art for automatic legal clause classification, there is no one-size-fits-all "best" approach for every situation. Practitioners have to weigh up accuracy, cost constraints, interpretability, and project goals when choosing a solution. Future research could include hybrid architectures, model distillation for saving resources, or innovative pretraining techniques to further enhance performance on a wide variety of legal document types. This comparative framework provides both a benchmark and a foundation for such future advances.

REFERENCES

[1] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Legal-BERT: The Muppets straight out of Law School," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, 2020. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.261

[2] I. Chalkidis *et al.*, "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," *arXiv preprint arXiv:2110.00976*, 2021. [Online]. Available: https://arxiv.org/abs/2110.00976

[3] S. M. H. Dadgar, M. S. Araghi, and M. R. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," in *Proc. IEEE Int. Conf. Eng. Technol. (ICETECH)*, pp. 112–116, 2016.

[4] X. L. Xiao, J. Wang, and S. Zuo, "Research on patent text classification based on Word2Vec and LSTM," in *Proc. IEEE Int. Conf. Big Data Analysis (ICBDA)*, pp. 1–6, 2019.

[5] I. Chalkidis, M. Fergadiotis, P. Androutsopoulos, and N. Aletras, "Neural contract element extraction using contextual embeddings," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 05, pp. 13976–13983, 2020.

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

[7] Y. Goldberg and O. Levy, "Word2Vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method," *arXiv preprint arXiv:1402.3722*, 2014.

[8] M. Boella, L. Di Caro, F. Grassi, A. Lopopolo, and L. Robaldo, "Linking legal open data: Breaking the accessibility and language barrier in European legislation and case law," in *Proc. Int. Conf. Artificial Intelligence and Law (ICAIL)*, pp. 171–175, 2018.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 10, pp. 3619–3639, 2018.

[12] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5753–5763, 2019.

[13] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2023.

[14] M. Kanapala, S. Pal, and R. Pamula, "Text summarization from legal documents: A survey," *Artificial Intelligence Review*, vol. 51, pp. 371–402, 2019.

[15] A. D. B. Carvalho, "Improving Legal BERT Models through Multi-Task Learning," *IEEE Access*, vol. 10, pp. 92205–92217, 2022.

[16] B. Liu, J. Hu, and J. Cheng, "Comparative analysis of Word2Vec and GloVe with LSTM for sentiment analysis," *Int. J. Scientific Technol. Res.*, vol. 10, no. 4, pp. 213–219, 2021.

[17] M. Qiu and S. Huang, "Text classification research based on TF-IDF and deep learning," *IEEE Access*, vol. 8, pp. 13545–13558, 2020.

[18] C. Cardellino, "Spanish Billion Words Corpus and Embeddings (Spanish Billion Words project)," 2016. [Online]. Available: http://crscardellino.me/SBWCE

[19] H. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[20] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," in *Proc. ACL*, pp. 328–339, 2018.

[21] S. Sweta "Modern Approach to Education Data mining and Its Applications", Published by Springer Nature- "Springer Briefs in Computational Intelligence"

[22] S. Sweta and K. Lal, "Adaptive e-Learning System: A State of Art," International Journal on Computer Application: ISSN 0975-8887, vol. 107, no. 7, pp. 13–15, 2014.