



Vidyavardhini's College of Engineering and Technology, Vasai

Department of Computer Science & Engineering (Data Science)

AY: 2025-26

<b>Class:</b>	<b>BE-CSE(DS)</b>	<b>Semester:</b>	<b>VII</b>
<b>Course Code:</b>	<b>CSDOL7011</b>	<b>Course Name:</b>	<b>NLP Lab</b>

<b>Name of Student:</b>	<b>Sahil Salunke</b>
<b>Roll No. :</b>	<b>45</b>
<b>Experiment No.:</b>	<b>1</b>
<b>Title of the Experiment:</b>	<b>Identifying and Critically Reviewing Research Papers on a Selected NLP Application</b>
<b>Date of Performance:</b>	
<b>Date of Submission:</b>	

### Evaluation

<b>Performance Indicator</b>	<b>Max. Marks</b>	<b>Marks Obtained</b>
Performance	5	
Understanding	5	
Journal work and timely submission	10	
<b>Total</b>	<b>20</b>	

<b>Performance Indicator</b>	<b>Exceed Expectations (EE)</b>	<b>Meet Expectations (ME)</b>	<b>Below Expectations (BE)</b>
Performance	4-5	2-3	1
Understanding	4-5	2-3	1
Journal work and timely submission	8-10	5-8	1-4

Checked by

**Name of Faculty** :

**Signature** :

**Date** :



Vidyavardhini's College of Engineering and Technology, Vasai  
Department of Computer Science & Engineering (Data Science)

**Aim:** To perform a critical literature review of five research papers on a chosen NLP application, focusing on the problem addressed, solution proposed, and limitations identified in each work.

**Objective:** To critically review five research papers on an NLP application, analyzing problems, solutions, and limitations.

**Tools Required:**

1. Research databases: Google Scholar, IEEE Xplore, ACM Digital Library, Springer, Elsevier, or arXiv
2. MS Word / Google Docs or LaTeX for documentation
3. Internet access

**Procedure:**

1. Select a Real-World NLP Application:
  - a. Choose any one topic such as:
    - i. Machine Translation
    - ii. Sentiment Analysis
    - iii. Text Summarization
    - iv. Question Answering Systems
    - v. Chatbots
    - vi. Named Entity Recognition
    - vii. Information Retrieval, etc.
2. Search and Select 5 Research Papers:
  - a. Preferably from peer-reviewed journals or conferences.
  - b. Papers must be recent (preferably from the last 5–7 years).
  - c. Ensure papers are directly relevant to the chosen application.
3. Read and Analyze Each Paper:
  - a. Focus on these three aspects for each paper:
    - i. Problem Statement: What problem or challenge does the paper address?
    - ii. Proposed Solution: What model, algorithm, or framework is presented?



- iii. Critical Evaluation: What are the limitations, gaps, or areas for improvement?
4. Prepare the Review Document:
  - a. Create a structured table or section for each paper with:
    - i. Title, Authors, Year, and Source
    - ii. Summary of problem
    - iii. Summary of solution
    - iv. Critical remarks and insights
5. Submit the Review:
  - a. The write-up should be minimum 4–5 pages.
  - b. Include references in standard citation format (APA/IEEE/MLA).

### **Description of the Experiment:**

This experiment introduces students to academic research and the process of reviewing scientific literature in the field of NLP. It allows them to explore cutting-edge developments, analyze technical approaches, and reflect critically on current limitations. This lays the foundation for their final-year projects or research internships.

### **Detailed Description of the NLP Technique:**

Since this experiment is open-ended, the NLP techniques will vary based on the papers selected. However, students will encounter:

- Deep learning architectures (e.g., LSTM, BERT, GPT, T5)
- Statistical NLP approaches (e.g., N-gram models, HMMs)
- Evaluation metrics (e.g., BLEU, ROUGE, Accuracy, F1-score)
- Datasets commonly used for training and evaluation

### **Conclusion:**



Name of the Paper	Solution Used	Drawbacks / Limitations	Areas of Improvement
<b>1]</b> Named Entity Recognition (NER) for Legal Document Analysis.	Used CRF, BiLSTM-CRF, and Transformer-based models (especially Legal-BERT) on Indian legal corpora. Achieved best performance with Legal-BERT (F1-score: 91.2%).	<ul style="list-style-type: none"><li>- Ambiguity in entities (e.g., "Delhi" as city/jurisdiction/organization)</li><li>- Lack of annotated corpora</li><li>- Multilingual complexity in Indian legal texts</li><li>- Ethical risks in automated legal AI</li></ul>	<ul style="list-style-type: none"><li>- Develop multilingual NER models</li><li>- Improve dynamic entity linking across jurisdictions</li><li>- Address fairness and bias in AI models</li><li>- Enhance corpus size and quality</li></ul>
<b>2]</b> Named Entity Recognition for Serbian Legal Documents.	Fine-tuned a Serbian-specific BERT model (BERTiĆ) for NER task on a manually annotated dataset of appellate court rulings. Achieved an average F1-score of 0.96.	<ul style="list-style-type: none"><li>- Small-scale dataset</li><li>- Cyrillic-to-Latin conversion required</li><li>- Class imbalance (high number of "O" tokens)</li><li>- Batch size limitations due to input length</li></ul>	<ul style="list-style-type: none"><li>- Expand the annotated dataset</li><li>- Develop more efficient training techniques</li><li>- Address token class imbalance</li><li>- Improve multilingual support in PTMs</li></ul>
<b>3]</b> Improving Legal Entity Recognition Using a Hybrid Transformer Model and Semantic Filtering Approach.	Combined Legal-BERT with a semantic similarity filtering step to refine predictions. Cosine similarity used to retain valid entities. Achieved 93.4% F1-score.	<ul style="list-style-type: none"><li>- High computational cost due to hybrid architecture</li><li>- Requires high-quality predefined legal patterns</li><li>- Potential dependency on domain-specific patterns</li></ul>	<ul style="list-style-type: none"><li>- Optimize computational efficiency</li><li>- Generalize filtering to broader legal subdomains</li><li>- Automate pattern generation for new domains</li></ul>
<b>4]</b> Extracting Complex Named Entities in Legal Documents via Weakly Supervised Object Detection.	Used weakly supervised object detection and Document Layout Analysis (DLA) with pseudo-label generation and a multi-modal approach (LayoutLMv3).	<ul style="list-style-type: none"><li>- Dependent on OCR accuracy</li><li>- Requires careful rule design for pseudo labeling</li><li>- Performance drop in highly unstructured layouts</li></ul>	<ul style="list-style-type: none"><li>- Improve OCR robustness</li><li>- Enhance object detection for irregular layouts</li><li>- Expand exact label availability for supervised training</li></ul>