**AY: 2025-26**

| **Class:** | **BE- CSE(DS)** | **Semester:** | **VII** |
| --- | --- | --- | --- |
| **Course Code:** | **CSDOL7011** | **Course Name:** | **NLP Lab** |

| **Name of Student:** | Soham Shivpuje |
| --- | --- |
| **Roll No. :** | 50 |
| **Experiment No.:** | **1** |
| **Title of the Experiment:** | **Identifying and Critically Reviewing Research Papers on a Selected NLP Application** |
| **Date of Performance:** |  |
| **Date of Submission:** |  |

# **Evaluation**

| **Performance Indicator** | **Max. Marks** | **Marks Obtained** |
| --- | --- | --- |
| Performance | 5 |  |
| Understanding | 5 |  |
| Journal work and timely submission | 10 |  |
| Total | 20 |  |

| **Performance Indicator** | **Exceed Expectations (EE)** | **Meet Expectations (ME)** | **Below Expectations (BE)** |
| --- | --- | --- | --- |
| 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 :**

**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:
   1. Choose any one topic such as:
      1. Machine Translation
      2. Sentiment Analysis
      3. Text Summarization
      4. Question Answering Systems
      5. Chatbots
      6. Named Entity Recognition
      7. Information Retrieval, etc.
2. Search and Select 5 Research Papers:
   1. Preferably from peer-reviewed journals or conferences.
   2. Papers must be recent (preferably from the last 5–7 years).
   3. Ensure papers are directly relevant to the chosen application.
3. Read and Analyze Each Paper:
   1. Focus on these three aspects for each paper:
      1. Problem Statement: What problem or challenge does the paper address?
      2. Proposed Solution: What model, algorithm, or framework is presented?
      3. Critical Evaluation: What are the limitations, gaps, or areas for improvement?
4. Prepare the Review Document:
   1. Create a structured table or section for each paper with:
      1. Title, Authors, Year, and Source
      2. Summary of problem
      3. Summary of solution
      4. Critical remarks and insights
5. Submit the Review:
   1. The write-up should be minimum 4–5 pages.
   2. 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** |
| --- | --- | --- | --- |
| **1) 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%). | - Ambiguity in entities (e.g., “Delhi” as city/jurisdiction/organization) - Lack of annotated corpora - Multilingual complexity in Indian legal texts - Ethical risks in automated legal AI | - Develop multilingual NER models - Improve dynamic entity linking across jurisdictions - Address fairness and bias in AI models - Enhance corpus size and quality |
| **2) 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. | - Small-scale dataset - Cyrillic-to-Latin conversion required - Class imbalance (high number of “O” tokens) - Batch size limitations due to input length | - Expand the annotated dataset - Develop more efficient training techniques - Address token class imbalance - Improve multilingual support in PLTMs |
| **3) Improving Legal Entity Recognition Using a Hybrid Transformer and Semantic Filtering Approach.** | Combined Legal-BERT with a semantic similarity filtering step to refine predictions. Some domain-specific patterns used to retain valid entities. Achieved 93.4% F1-score. | - High computational cost due to hybrid architecture - Requires high-quality predefined legal patterns - Potential dependency on domain-specific patterns | - Optimize computational efficiency - Generalize filtering to broader legal subdomains - Automate pattern generation for new domains |
| **4) Extracting Entities from Complex Scanned Legal Documents Using a Weakly Supervised Framework with Document Layouts and Object Detection.** | Used weakly supervised object detection and Document Layout Analysis (DLA) with pseudo-labeling in visually unstructured documents. (LayoutLMv3). | - Dependent on OCR accuracy - Requires careful rule design for pseudo labeling - Performance drop in highly unstructured layouts | - Improve OCR robustness - Enhance object detection for irregular layout - Use few-shot or self-supervised learning techniques |
| **5) Named Entity Recognition on Indonesian Legal Documents.** | Constructed “IndoLER” dataset (~1,000 annotated court documents) with 20 fine-grained legal entity types. Fine-tuned transformer models (XLM-RoBERTa, IndoRoBERTa, etc.) on the dataset. | - Moderate corpus size limits generalization - Limited jurisdictional applicability - Computational demands of fine-tuning | - Expand dataset size and legal domains - Support multilingual or low-resource jurisdictions - Reduce class imbalance - Enable transfer learning across legal systems |